University of Cape Town

Department of Statistical Sciences

VARIABLE SELECTION IN LOGISTIC REGRESSION, WITH SPECIAL APPLICATION TO MEDICAL DATA

GEORGINA JOUBERT

A dissertation prepared in fulfilment of the requirements for the degree of Master of Science in Mathematical Statistics

Supervisor: Prof Walter Zucchini

November 1994

ACKNOWLEDGEMENTS

I thank

my supervisor, Prof. Walter Zucchini, who gave me constant support and encouragement, and remained patient through the long years

my husband, Robert Schall, who was less patient, but gave valuable support and guidance during the write-up of the dissertation

Prof. Willie Mollentze of the Department of Internal Medicine, University of the Orange Free State; and the Mamre Community Health Project of the University of Cape Town and the Medical Research Council, for approaching me to do the analysis of their challenging data sets.

TABLE OF CONTENTS

LIST OF TABLES

ACKNOWLEDGEMENTS

I thank

my supervisor, Prof. Walter Zucchini, who gave me constant support and encouragement, and remained patient through the long years

my husband, Robert Schall, who was less patient, but gave valuable support and guidance during the write-up of the dissertation

Prof. Willie Mollentze of the Department of Internal Medicine, University of the Orange Free State; and the Mamre Community Health Project of the University of Cape Town and the Medical Research Council, for approaching me to do the analysis of their challenging data sets.

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER 1: INTRODUCTION

1.1 Binary outcome

In medical data the outcome of interest is often dichotomous: the patient is alive or dead, the patient recovered or did not recover after a certain time interval, the individual is diseased or not diseased. Although the outcome variable may in fact be measured on a continuous scale (for example blood pressure in mm Hg) the clinician may be more interested in a dichotomised outcome (for example hypertensive and not hypertensive).

1.2 Factors associated with the outcome

Rather than just describing how many patients are alive or dead, diseased or not diseased, the researcher may wish to know whether the outcome is associated with other characteristics (for example, is a higher proportion of women/older people/overweight people hypertensive?). There may be many such factors, some of which are dichotomous, some categorical and some continuous.

1.2.1 Assessment of association between a single dichotomous factor and the outcome

If there is just one dichotomous characteristic (say, exposure) which is to be related to the dichotomous outcome (say, disease), Fisher's exact test or a chi-squared test can be performed to assess the statistical significance of the association between exposure and outcome in the following 2x2 table (Fleiss 1981):

		OUTCOME			
		Diseased	Not diseased		
F					
A	Factor present	a	b	a+b	(N_1)
C	(Exposed)				
T	Factor absent	c	d	c+d	(N_2)
O	(Not exposed)				
R					
		a+c	b+d	a+b+c+d	(N)
		(M_1)	(M_2)		

From this table measures of the strength of the association can be calculated. Statistical measures of the strength of association between a factor (exposure) and outcome (disease) measure the way in which the risks of disease differ among those with the factor present (the exposed) and those without the factor (the non-exposed). Two such measures are the relative risk (risk ratio) and the odds ratio (Fleiss 1981). The relative risk is the ratio of the risk of disease among the exposed (R_1) and the risk of disease among the non-exposed (R_2), ie R_1/R_2. If R is the risk of disease then the odds of disease are $R/(1-R)$. The odds ratio (OR) is the ratio of the odds of disease among the exposed, $R_1/(1-R_1)$, and the odds of disease among the non-exposed, $R_2/(1-R_2)$:

$$OR = \frac{R_1/(1-R_1)}{R_2/(1-R_2)}$$

Since risks of disease are probabilities and lie between 0 and 1, the relative risk and odds ratio are positive numbers. A relative risk or odds ratio of 1 or close to 1 indicates that the exposure is neither a risk factor nor a protective factor, there is no association between the exposure and the outcome. A value larger than 1 indicates that the exposure is a risk

factor, a value smaller than 1 that the exposure is a preventative factor. Since the relative risk and odds ratio calculated in a study are subject to sampling variation, the estimates should be reported and interpreted with their confidence intervals. Whereas hypothesis testing assesses the statistical significance of the association, the estimates and confidence intervals for the relative risk and odds ratio can be used to evaluate the clinical significance of an association.

If the independent factor is a categorical variable with k levels ($k > 2$), relative risks and odds ratios can be calculated as above, by choosing one of the levels as a reference group and comparing each of the other levels, in turn, with the reference group. A continuous variable can be categorised and the above approach can then be used.

1.2.2 Study design

Information about a factor and an outcome can be collected in different ways. Analytical studies in which exposure and disease information is collected on a sample of subjects can be performed in three different ways, which are distinguished by the sampling scheme used.

In a cross-sectional study a sample of N subjects is selected, and then cross-classified with respect to exposure and disease. In a follow-up study (also called cohort or prospective study) a sample of N_1 subjects with the exposure and a sample of N_2 subjects without the exposure are selected and followed up to investigate who develops the disease and who does not. The subjects in a follow-up study are thus sampled according to exposure status. In a case-control (or retrospective) study subjects are sampled according to disease outcome: a sample of M_1 cases with the disease and M_2 controls without the disease are selected and retrospective information on exposure is collected to determine what proportion of each group had been exposed. For a rare disease the case-control study is often the only way in which the association between exposure and disease can be investigated, since in a

cross-sectional study few subjects with the rare disease will be sampled, and in a follow-up study few subjects would develop the disease.

Experimental studies differ from analytical studies in that the researcher in an experimental study does not rely on observing but experiments actively. In a clinical trial subjects (patients or volunteers) are randomised to different treatments and followed up to determine outcome (for example recovery, or the presence of adverse events).

1.2.3 Study design and measures of association

The type of study design determines which measures of the strength of association can be estimated. In the case of a cross-sectional, follow-up study or experimental study (clinical trial) the risk of disease in those with the factor can be estimated from the 2x2-table by $a/(a+b)$, and the risk of disease in those without the factor as $c/(c+d)$. The relative risk can then be estimated by $a(c+d)/c(a+b)$, which is a measure of the strength of association between the factor and the disease.

In a case-control study, however, a predetermined number of cases (diseased) and controls (not diseased) are sampled and their numbers generally do not reflect the proportions of diseased and non-diseased in the population. It is therefore not possible to estimate the risk of disease among those with and without the risk factor so that the relative risk cannot be calculated.

In cross-sectional, follow-up and experimental studies, the odds of disease can be defined as the probability of disease compared to the probability of no disease. So, in the group with the factor present, the odds of disease is estimated by $a/(a+b) \div b/(a+b) = a/b$. Similarly the odds of disease in those without the factor is estimated by c/d. The odds ratio is then the odds of disease in those with the factor, compared to the odds of disease in those without the factor, which can be estimated by $a/b \div c/d = ad/bc$.

In a case-control study it is not possible to estimate the odds of disease among those with and without the risk factor. A case-control study does,

however, provide information on the odds of exposure among the diseased (a/c) and not diseased (b/d), which can be used to estimate the odds ratio as $a/c \div b/d = ad/bc$. The odds ratio can thus be estimated for all epidemiologic study designs. It is because of this that the odds ratio has gained much popularity as a measure of the strength of association between an exposure and an outcome variable, in particular in case-control studies. The odds ratio and its $100(1-\alpha)\%$ confidence interval are therefore often used to assess the strength of the association.

For a rare disease (generally studied by means of a case-control study) the odds ratio approximates the relative risk, since a and c are small, and thus

$$\frac{a}{a+b}\frac{c+d}{c} \approx \frac{a}{b}\frac{d}{c}.$$

This is another reason for the popularity of the odds ratio as measure of the strength of association, and for the use of case-control studies.

1.2.4 Assessing the association between more than one factor and the outcome

The measures of strength of association outlined above were used to assess the association between one factor (exposure) and the outcome. However, in most studies several characteristics of the patient or individual (physical, treatment, exposure to risk factors) are measured, and their association with the outcome are to be assessed.

Example 1.1

In a cross-sectional community survey in Mangaung, an urban township in the Orange Free State, a sample of 758 adults (aged 25 years and older) was studied to determine the prevalence of factors related to cardiovascular disease (Mollentze et al 1994). So, for example, data were collected on sex, blood pressure and hypertension treatment (individuals were thus classified as hypertensive or not), glucose levels and diabetic treatment (individuals were thus classified as having impaired glucose tolerance or not).

The outcome of interest is glucose tolerance status, the exposure of interest hypertension status. In the over 45 year old group the percentages of individuals with impaired glucose tolerance, by sex and hypertension status, were as follows:

male hypertensives 22/67=33%

male normotensives 13/88=15%

female hypertensives 52/144=36%

female normotensives 28/85=33%

To compare hypertensives with normotensives on the outcome impaired glucose tolerance one could ignore the variable sex and analyse the association as outlined above. However, of the hypertensives 144/211=68% are females whereas only 85/173=49% of the normotensives are females. In comparing the hypertensives with the normotensives one would therefore wish to take the variable sex into account in some way.

1.2.5 Stratified analysis

By using a stratified analysis one attempts to adjust for, or remove, the effect of an extraneous variable. So, in Example 1.1, one could stratify the data by sex and assess the association between hypertension and glucose tolerance for each sex separately. If the odds ratios in the different strata are similar, and similar to the unstratified (crude) odds ratio, the crude odds ratio can be reported. If the odds ratios in the different strata are similar, but different from the crude odds ratio, an adjusted (common) odds ratio can be calculated using the Mantel-Haenszel approach, which is a weighted average of the stratum-specific odds ratios (Mantel and Haenszel 1959). If the odds ratios in the different strata differ markedly the stratum-specific odds ratios should be reported. This type of analysis is mostly done in epidemiology where the aim of the analysis is estimation: the researcher wishes to assess the association between an exposure and an outcome, while adjusting for other variables. In epidemiology analytical studies (also called observational

studies) are mostly used, and the groups to be compared can differ vastly with respect to other variables (commonly sex and age). In clinical trials, because of randomisation of subjects to treatment groups (which can also take the form of stratified randomisation) other variables may not be as important.

1.2.6 Logistic regression

To adjust for a factor by means of a stratified analysis continuous variables need to be categorised. However, stratification becomes inefficient if there are many independent variables to consider (in Example 1.1, apart from sex, one might wish to take age and obesity into account as well). In such a case a logistic regression model which is a statistical model which describes the relationship between the outcome and independent variables (which can be continuous) can be fitted. As an example, Truett, Cornfield and Kannel (1967) describe the usefulness of the logistic model for the community-based follow-up study in Framingham.

1.3 Uses of logistic regression in medical research

Logistic regression in medical research is used for two different aims: on the one hand it is used to select important predictors of outcome, or to form a prediction equation to predict outcome of future observations. On the other hand, it is used to investigate the effect of a certain factor (exposure) on an outcome (commonly measured through the odds ratio) while adjusting for other factors (confounders). A mixture of these two aims can also be encountered: the researcher may wish to identify important predictors, while adjusting for certain factors (commonly age or sex). A review of articles published in the American Journal of Public Health in 1992 showed that in 17 articles and 17 briefs (short reports) logistic regression was used to identify predictors. In 14 articles and 16 briefs the logistic regression was used for estimation, adjusting for confounders, and in 5 articles and 1 brief the aim of the logistic regression was a mixture of prediction and estimation.

The following two examples illustrate the type of logistic regression seen in the medical context.

Example 1.2

A follow-up study was done on a group of head injured children (55 moderately, 40 severely and 28 very severely head injured) and a group of 46 controls (children hospitalised for arm and leg fractures but with no head injury) (Hemp 1989). At admission to hospital (the intake phase) a medical questionnaire was completed describing the details of the injury and accident (for example, pedestrian or fall, skull fractured, type of coma, post traumatic amnesia). A social questionnaire was also completed by questioning the parents/guardians (for example, social background of family - income, overcrowding-, social behaviour and school/preschool background of child). Shortly after the injury a battery of neuropsychological tests were performed, including IQ tests and tests to evaluate language and motor skills. These tests were repeated after three months, and again at approximately one year after the injury. Apart from describing recovery patterns over the year in the four different groups, the aim of the project was to form a prediction equation based on intake variables (medical, social and neuro-psychological) in the head injured group to predict outcome at one year after injury.

Example 1.3

There is currently much interest in the clinical medical literature on the relationship between hypertension and hyperinsulinaemia. A study was conducted on a sample of 854 QwaQwa residents aged 25 years and older (Mollentze et al 1994). Amongst others, information was gathered on factors associated with hypertension and hyperinsulinaemia. Obesity, age, sex and diabetes are known to be related to both hypertension and hyperinsulinaemia, with obese people, older people and diabetics being more likely to be hypertensive as well as hyperinsulinaemic. To investigate the relationship between hypertension and hyperinsulinaemia it is therefore necessary to

consider controlling for these variables in some way.

1.4 Variable selection

As the two examples in the previous section indicate, there is often a large pool of possible predictors to select from, or possible factors to adjust for, in medical applications of logistic regression. Fitting all possible predictors in a prediction problem could lead to an equation which is not easily generalisable nor of much practical use. Adjusting for all possible confounders in an estimation problem may reduce the chance of biased estimates but could lead to numerically unstable or highly variable estimates. The choice of which variables to include in the model is thus central to most logistic regression analyses.

For both prediction and estimation problems various selection procedures have been proposed in the statistical, epidemiological and medical literature. The procedures can be broadly categorised as

- hypothesis testing
- the comparison of various models on the basis of some criterion, where the criterion chosen depends on whether the aim of the logistic regression is estimation or prediction, or
- procedures which place emphasis on the known biological importance of variables, and biologically expected sizes of effects.

1.5 Aim of this thesis

In this thesis the various methods of variable selection which have been proposed in the statistical, epidemiological and medical literature for prediction and estimation problems in logistic regression will be described. The procedures will be applied to medical data sets. On the basis of the literature review as well as the applications to examples, strengths and weaknesses of the approaches will be identified. The procedures will be compared on the basis of the results obtained, their appropriateness for the specific aim of the

analysis, and demands they place on the analyst and researcher, intellectually and computationally. In particular, certain selection procedures using bootstrap samples, which have not been used before, will be investigated, and the partial Gauss discrepancy will be extended to the case of logistic regression. Recommendations will be made as to which approaches are the most suitable or most practical in different situations. Most statistical texts deal with issues regarding prediction, whereas the epidemiological literature focus on estimation. It is therefore hoped that the thesis will be a useful reference for those, statistically or epidemiologically trained, who have to deal with issues regarding variable selection in logistic regression.

When fitting models in general, and logistic regression models in particular, it is standard practice to determine the goodness of fit of models, and to ascertain whether outliers or influential observations are present in a data set. These aspects will not be discussed in this thesis, although they were considered when fitting the models.

CHAPTER 2: THE LOGISTIC REGRESSION MODEL

In this chapter the multiple linear logistic regression model will be described, as well as the methods employed to obtain estimates of the coefficients. The deviance, which is often used to decide whether further terms need to be included in a model, will be discussed. The interpretation of the coefficients in the logistic model will be outlined and the rationale for variable selection will be given.

2.1 Binary outcome and the Binomial distribution

Suppose that the outcome of the individuals under investigation is classified into one of two categories, say "success" and "failure", where these two generic terms represent the outcomes of interest, for example alive/dead, recovered/not recovered, diseased/not diseased. Let y denote the outcome of a given individual.

$$y = 1 \text{ if the outcome is a success}$$

$$y = 0 \text{ if the outcome is a failure}$$

Let $\pi = P(y = 1)$ and so $P(y = 0) = 1 - \pi$. Suppose also that data on p (say) predictor variables are available for each individual, $x_1, ..x_p$. The objective of many investigations is to examine the relationship between π and the predictor variables.

The outcome of the i-th individual can also be viewed as a proportion y_i/n_i where

$$n_i = 1 \qquad i = 1, \ldots, n$$

$$y_i = 1 \text{ if outcome=success}$$

$$y_i = 0 \text{ if outcome=failure.}$$

This is a special case of so-called grouped data, where the observations are of the form y_i/n_i with y_i the number of successes out of n_i trials.

The appropriate distribution for y_i is $B(1, \pi_i)$. Using the properties of the Binomial distribution

$$E(y_i) = \pi_i$$

$$\text{Var}(y_i) = \pi_i(1 - \pi_i)$$

2.2 The logistic regression model

In linear regression with a continuous outcome variable y_i, a model of the form

$$E(y_i) = \mu_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_p x_{pi} \qquad i = 1, \ldots, n$$

could be fitted (Draper and Smith 1981). So, for example, a linear regression model could be fitted to describe blood pressure measured in mm Hg by means of the predictor variables age, sex and body mass index. However, blood pressure and information on the use of antihypertensive treatment could be used to categorise individuals as hypertensive or not hypertensive, and the clinician may be interested in a model which describes hypertensive state by means of age, sex and body mass index.

There are, however, problems with expressing $E(y_i) = \pi_i$ as a linear combination of the predictor variables: π_i lies between 0 and 1 whereas a linear combination of the predictor variables could lead to an outcome in the range $(-\infty, \infty)$. This problem can be overcome by employing a transformation (g) of π_i that maps (0,1) onto $(-\infty, \infty)$. Then $g(\pi_i)$ can be written as a linear combination of the predictor variables

$$g(\pi_i) = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j \qquad i = 1, \ldots, n$$

One such transformation is the logistic transformation, the logit, (McCullagh and Nelder 1989)

$$g(\pi_i) = \ln\frac{\pi_i}{1 - \pi_i} = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j \qquad i = 1, \ldots, n$$
$$= \eta_i$$

This model can be used to incorporate transformed predictor variables (for example $x_1 = \ln(age)$), as well as products of predictor variables. Furthermore, a variable with k categories ($k > 2$) can be considered by creating $(k-1)$ design variables and including these $(k-1)$ design variables as a set in the model. For example, if a variable has 3 categories (residence is rural, town or city), two design variables can be created as follows

	design variable 1	design variable 2
category 1	0	0
category 2	1	0
category 3	0	1

2.3 Estimation (McCullagh and Nelder 1989)

2.3.1 Likelihood function

The coefficients β_0 to β_p can be estimated by maximising the likelihood function. For individuals with $y_i = 0$, the contribution to the likelihood is $(1-\pi_i)^{1-y_i}$. If $y_i = 1$ the contribution is $\pi_i^{y_i}$. The contribution of any observation is thus $\pi_i^{y_i}(1-\pi_i)^{1-y_i}$. The likelihood function for n observations is thus

$$L(\beta) = \prod_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{1-y_i} \qquad i = 1, \ldots, n \tag{2.1}$$

The likelihood depends on the π_i which in turn depend on β.

Maximum likelihood estimates $\hat{\beta}_0, \hat{\beta}_1, \ldots \hat{\beta}_p$ are obtained by finding the

values which maximise L, or equivalently $\ln L$.

$$\begin{aligned}
\ln L(\beta) &= \sum_{i=1}^{n}[y_i \ln \pi_i + (1-y_i)\ln(1-\pi_i)] \\
&= \sum_{i=1}^{n}[y_i \ln \frac{\pi_i}{1-\pi_i} + \ln(1-\pi_i)] \\
&= \sum_{i=1}^{n}[y_i \eta_i + \ln(1 - \frac{e^{\eta_i}}{1+e^{\eta_i}})] \\
&= \sum_{i=1}^{n}[y_i \eta_i - \ln(1+e^{\eta_i})]
\end{aligned}$$

where $\eta_i = \ln \frac{\pi_i}{1-\pi_i} = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ji}$

Taking partial derivatives

$$\frac{\partial \ln L}{\partial \beta_0} = \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} e^{\eta_i}(1+e^{\eta_i})^{-1}$$

$$\frac{\partial \ln L}{\partial \beta_j} = \sum_{i=1}^{n} y_i x_{ji} - \sum_{i=1}^{n} e^{\eta_i}(1+e^{\eta_i})^{-1} x_{ji} \qquad j = 1, \ldots, p$$

Equating these $p+1$ equations to zero gives a set of $p+1$ non-linear equations which have to be solved.

2.3.2 Newton-Raphson procedure

The score of the j-th parameter is $\partial \ln L / \partial \beta_j$. Denote the $(p+1) \times 1$ vector of scores by $U(\beta)$. A $(p+1) \times (p+1)$ matrix of second order partial derivatives can be formed with (i,j)-th element

$$\frac{\partial^2 \ln L}{\partial \beta_i \partial \beta_j} \qquad i = 0, \ldots, p \qquad j = 0, \ldots, p$$

This matrix is called the Hessian matrix, denoted by $H(\beta)$. Near $\hat{\beta}$, at β_m, the Taylor's expansion of the scores vector gives

$$U(\hat{\beta}) \approx U(\beta_m) + H(\beta_m)(\hat{\beta} - \beta_m) \tag{2.2}$$

The maximum likelihood estimates of the $\beta's$ must satisfy

$$\frac{\partial \ln L}{\partial \hat{\beta}_j} = 0$$

so

$$U(\hat{\beta}) = 0$$

and from (2.2)

$$U(\beta_m) + H(\beta_m)(\hat{\beta} - \beta_m) = 0$$

Thus

$$\hat{\beta} \approx \beta_m - H^{-1}(\beta_m)U(\beta_m)$$

which suggests an iterative scheme for estimating $\hat{\beta}$

$$\hat{\beta}_{m+1} = \hat{\beta}_m - H^{-1}(\hat{\beta}_m)U(\hat{\beta}_m) \tag{2.3}$$

2.3.3 Fisher scoring

An alternative method of solving the likelihood equations is the Fisher scoring method. In this method the Hessian matrix is replaced by the matrix of expected values of second order partial derivatives. The information matrix I has (j,k)-th element

$$-E[\frac{\partial^2 L}{\partial\beta_j \partial\beta_k}].$$

The iterative scheme is then

$$\hat{\beta}_{m+1} = \hat{\beta}_m + I^{-1}(\hat{\beta}_m)U(\hat{\beta}_m) \tag{2.4}$$

In the case of the linear logistic model $I^{-1}(\hat{\beta}) = -H^{-1}(\hat{\beta})$ so the two algorithms will not only converge to the maximum likelihood estimate of β, but will give the same standard errors of the parameter estimates (the square root of the diagonal elements of $-H^{-1}$ and I^{-1}).

2.3.4 Iteratively reweighted least squares

To fit the linear logistic model using Fisher scoring, we need expressions for $U(\beta)$ and $I(\beta)$. From (2.1) the log-likelihood function for n observations is given by

$$\ln L = \sum_{i=1}^{n} [y_i \ln \pi_i + (1 - y_i)\ln(1 - \pi_i)]$$

Thus

$$\frac{\partial \ln L}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial \ln L}{\partial \pi_i} \frac{\partial \pi_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_j}$$

The three components are as follows

$$\begin{aligned} \frac{\partial \ln L}{\partial \pi_i} &= \sum_{i=1}^{n} \frac{y_i}{\pi_i} - \frac{1-y_i}{1-\pi_i} \\ &= \sum_{i=1}^{n} \frac{y_i - \pi_i}{\pi_i(1-\pi_i)} \qquad (2.5) \\ \frac{\partial \eta_i}{\partial \beta_j} &= x_{ji} \qquad (2.6) \\ \frac{\partial \pi_i}{\partial \eta_i} &= \frac{\partial \eta_i}{\partial \pi_i}^{-1} = g'(\pi_i) \\ &= \frac{1}{\pi_i(1-\pi_i)} \qquad (2.7) \end{aligned}$$

Therefore, from (2.5), (2.6) and (2.7)

$$\frac{\partial \ln L}{\partial \beta_j} = \sum_{i=1}^{n} \frac{y_i - \pi_i}{\pi_i(1-\pi_i)} \frac{1}{g'(\pi_i)} x_{ji}.$$

If $e_i = (y_i - \pi_i)g'(\pi_i)$ and

$$w_i = \frac{1}{\pi_i(1-\pi_i)[g'(\pi_i)]^2} = \pi_i(1-\pi_i)$$

then

$$\frac{\partial \ln L}{\partial \beta_j} = \sum x_{ji} w_i e_i$$

Therefore

$$U(\beta) = X'We$$

where X is the $n \times (p+1)$ matrix of the p predictor variables plus the intercept term, W the $n \times n$ diagonal matrix of weights w_i, and e is the $n \times 1$ vector with i-th component e_i. To obtain the (j,k)th element of the information matrix we use

$$-E(\frac{\partial^2 \ln L}{\partial \beta_j \partial \beta_k}) = E(\frac{\partial \ln L}{\partial \beta_j} \frac{\partial \ln L}{\partial \beta_k})$$

If $i \neq i'$ then

$$E[(y_i - \pi_i)(y_{i'} - \pi_{i'})] = Cov(y_i, y_{i'}) = 0$$

If $i = i'$ then

$$E[(y_i - \pi_i)^2] = \text{Var}(y_i) = \pi_i(1 - \pi_i)$$

Therefore

$$\begin{aligned} -E(\frac{\partial^2 \ln L}{\partial \beta_k \partial \beta_j}) &= \sum_{i=1}^{n} \frac{\pi_i(1-\pi_i)}{[\pi_i(1-\pi_i)]^2[g'(\pi_i)]^2} x_{ji}x_{ki} \\ &= \sum_{i=1}^{n} \frac{1}{\pi_i(1-\pi_i)} \frac{1}{[g'(\pi_i)]^2} x_{ji}x_{ki} \\ &= \sum_{i=1}^{n} x_{ji}w_i x_{ki} \end{aligned}$$

so that

$$I(\beta) = X'WX.$$

From (2.4) $\hat{\beta}_{m+1} = \hat{\beta}_m + I^{-1}(\hat{\beta}_m)U(\hat{\beta}_m)$ Thus

$$\begin{aligned} \hat{\beta}_{m+1} &= \hat{\beta}_m + (X'W_mX)^{-1}X'W_me_m \\ &= (X'W_mX)^{-1}[X'W_m(X\hat{\beta}_m + e_m)] \\ &= (X'W_mX)^{-1}X'W_m(\hat{\eta} + e) \\ &= (X'W_mX)^{-1}X'W_mz_m \end{aligned}$$

where subscript m denotes the values are obtained from the m-th iteration. $\hat{\beta}_{m+1}$ is thus obtained by regressing (using weighted least squares) the adjusted dependent variable z_m, with i-th element

$$\hat{\eta}_{im} + (y_i - \pi_i)g'(\pi_i) = \hat{\eta}_{im} + \frac{(y_i - \pi_i)}{\pi_i(1-\pi_i)}$$

on the p predictor variables using weights w_{im} where

$$w_{im} = \frac{1}{\pi_i(1-\pi_i)[g'(\pi_i)]^2} = \pi_i(1-\pi_i)$$

As initial estimates of π_i, $\hat{\pi}_{i0} = (y_i + 1/2)/2$ can be used (Collett 1991), from which initial values for the weights and adjusted dependent variable can be calculated. By performing weighted least squares regression on the adjusted dependent variable, estimates of $\hat{\beta}$ are obtained which lead to revised estimates of $\hat{\eta}$ and $\hat{\pi}$, the weights and adjusted dependent variable. The deviance (see section 2.4) is used to decide whether iteration should stop.

2.4 The deviance (McCullagh and Nelder 1989)

To measure the goodness of fit of a given model, one can consider comparing the value of the likelihood of the model, L_c, when the parameters are set equal to their maximum likelihood estimates to that of the full (saturated) model, L_f, where the full model is the model for which the fitted values coincide with the observed data. The deviance is defined as

$$\begin{aligned} D(c,f) &= -2\ln(\frac{L_c}{L_f}) \\ &= -2\ln L_c + 2\ln L_f. \end{aligned}$$

In the case of binary data

$$\begin{aligned} \ln L_c &= \sum_{i=1}^{n}[y_i \ln\hat{\pi}_i + (1-y_i)\ln(1-\hat{\pi}_i)] \\ \ln L_f &= \sum_{i=1}^{n}[y_i \ln y_i + (1-y_i)\ln(1-y_i)] \\ &= 0 \end{aligned}$$

Thus

$$\begin{aligned} D(c,f) &= -2\sum_{i=1}^{n}[y_i \ln\hat{\pi}_i + (1-y_i)\ln(1-\hat{\pi}_i)] \\ &= -2\sum_{i=1}^{n}[y_i \ln\frac{\hat{\pi}_i}{1-\hat{\pi}_i} + \ln(1-\hat{\pi}_i)] \\ &= -2\hat{\beta}'X'Y - 2\sum_{i=1}^{n}\ln(1-\hat{\pi}_i) \\ &= -2\hat{\eta}'\hat{\pi} - 2\sum_{i=1}^{n}\ln(1-\hat{\pi}_i) \end{aligned}$$

The deviance is therefore a function of $\hat{\pi}_i$ and $\hat{\beta}$ only, and conditional on the estimates of β the deviance does not depend on the data. Therefore the deviance can give no information about the fit of the model. Furthermore, in the case of $n_i = 1$ for $i = 1, \ldots, n$ the large sample approximation that $D \sim \chi^2_{n-p}$ does not hold (McCullagh and Nelder 1989).

However, the deviance can be used to compare two nested models. If $D_1 = -2[\ln L_{c1} - \ln L_f]$ and $D_2 = -2[\ln L_{c2} - \ln L_f]$ where model c_1 is nested within model c_2 (say c_1 contains p parameters, and model c_2 contains a further q parameters) then $D_1 - D_2 = -2[\ln L_{c1} - \ln L_{c2}]$ has an approximate χ^2_q-distribution (McCullagh and Nelder 1989).

2.5 Interpreting the coefficients

As described before, measures of the strength of association are useful for describing the relationship between exposure and outcome. Since the logistic regression models the log odds ratio of success, it enables the researcher to calculate odds ratios related to various variables, using the coefficients from the model (for all types of epidemiologic studies). A few situations will be outlined below.

2.5.1 One dichotomous variable

If the independent variable is dichotomous and coded as 1=present, 0=absent, we have that

$$\text{if } x = 1 \text{ and } y = 1 \text{ then } \pi = \frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}}$$

$$\text{if } x = 1 \text{ and } y = 0 \text{ then } 1-\pi = \frac{1}{1+e^{\beta_0+\beta_1}}$$

The odds of disease if $x = 1$ is $e^{\beta_0+\beta 1}$

$$\text{if } x = 0 \text{ and } y = 1 \text{ then } \pi = \frac{e^{\beta_0}}{1+e^{\beta_0}}$$

$$\text{if } x = 0 \text{ and } y = 0 \text{ then } 1-\pi = \frac{1}{1+e^{\beta_0}}$$

The odds of disease if $x = 0$ is e^{β_0}. The odds ratio is thus

$$\frac{e^{\beta_0+\beta_1}}{e^{\beta_0}} = e^{\beta_1}.$$

An approximate $100(1-\alpha)\%$ confidence interval can be obtained from the limits for β_1. If the $100(1-\alpha)\%$ confidence interval for β_1 is given by

$$\hat{\beta}_1 \pm z_{1-\alpha/2}\hat{SE}(\hat{\beta}_1)$$

the approximate confidence interval for the odds ratio is given by:

$$e^{\hat{\beta}_1 \pm z_{1-\alpha/2}\hat{SE}(\hat{\beta}_1)}.$$

Fleiss (1979) describes available exact and approximate confidence intervals for coefficients.

If the dichotomous variable is, however, coded as 1=present, -1=absent, the estimated odds ratio is

$$\frac{e^{\beta_0+\beta_1}}{e^{\beta_0-\beta_1}} = e^{2\beta_1}$$

and the approximate $100(1-\alpha)\%$ confidence interval is

$$e^{2\hat{\beta}_1 \pm 2z_{1-\alpha/2}\hat{SE}(\hat{\beta}_1)}$$

2.5.2 One variable with k categories

If the independent variable has k categorical levels, a set of $k-1$ design variables have to be formed. If there are 3 levels, 2 design variables are created as follows, using "reference cell coding" (Hosmer and Lemeshow 1989) or the partial method (BMDP 1983)

	design variable 1	design variable 2
level 1	0	0
level 2	1	0
level 3	0	1

Level 1 is chosen as reference group, and the odds of disease in each of the other levels can be expressed relative to the odds of disease in the

reference level. Calculating the odds of disease in each level as outlined before, the odds ratio, level 2 to level 1, is e^{β_1}, and for level 3 to level 1 e^{β_2}. An approximate confidence interval can be calculated as before. If odds ratios consisting of the odds of one of the levels compared to the odds of one of the other levels, not the reference level, are to be calculated, the odds ratio will be the exponential of the difference between the coefficients in question, and methods to calculate the confidence interval will be as below.

The three levels can however also be coded using the "deviation from means" coding (Hosmer and Lemeshow 1989) or the marginal method (BMDP 1983)

	design variable 1	design variable 2
level 1	-1	-1
level 2	1	0
level 3	0	1

The odds of disease in levels 1 and 2 are then, respectively, $e^{\beta_0-\beta_1-\beta_2}$ and $e^{\beta_0+\beta_1}$. The odds ratio (level 2 compared to level 1) is thus $e^{2\beta_1+\beta_2}$.

To calculate a confidence interval for the odds ratio, one needs to estimate the variance of the sum of the coefficients $(2\beta_1 + \beta_2)$. The variance of the log odds ratio is thus $4\text{Var}(\hat{\beta}_1) + 4\text{Cov}(\hat{\beta}_1, \hat{\beta}_2) + \text{Var}(\hat{\beta}_2)$. The standard error is the square root of this, and the approximate confidence interval is obtained by exponentiating the limits of the confidence interval for $2\beta_1 + \beta_2$.

2.5.3 One continuous variable

In the case of a continuous variable the odds ratio associated with a unit change in x_1 is

$$\frac{e^{\beta_0+\beta_1 x_1+\beta_1}}{e^{\beta_0+\beta_1 x_1}} = e^{\beta_1}.$$

It may be more meaningful to express the odds ratio associated with c units of change : $e^{c\beta_1}$. The odds ratio associated with a change of c units does not

depend on the value of x. An approximate confidence interval is calculated as before.

2.5.4 One dichotomous variable and one continuous variable

A logistic regression model is mostly employed when there is more than one explanatory variable to take into account. The model statistically adjusts for the presence of other factors, so that the association of one variable with the outcome can be investigated in such a way that one is sure the association is due to that variable and not because of differences in the distribution of other variables. Let us consider the case of one dichotomous variable (x_1) coded as present=1, absent=0, and one continuous variable (x_2). If one wants to investigate the effect of x_1 in the presence of x_2 one can calculate the odds of disease if $x_1 = 1$ and $x_2 = \bar{x}_2$ as $e^{\beta_0+\beta_1+\beta_2\bar{x}_2}$, and the odds of disease if $x_1 = 0$ and $x_2 = \bar{x}_2$ as $e^{\beta_0+\beta_2\bar{x}_2}$. The odds ratio is thus e^{β_1}.

If, on the other hand, the interest is focussed on the odds ratio associated with an increase of c units in x_2, one can calculate the odds of disease if $x_1 = 1$ and $x_2 = x_2 + c$ as $e^{\beta_0+\beta_1+\beta_2(x_2+c)}$, and the odds of disease if $x_1 = 1$ and $x_2 = x_2$ as $e^{\beta_0+\beta_1+\beta_2}$. The odds ratio associated with an increase of c units in x_2 is thus $e^{c\beta_2}$. Confidence intervals are calculated as before.

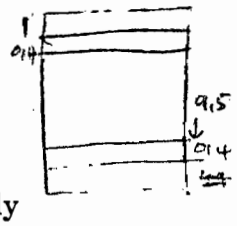
2.5.5 One dichotomous variable, one continuous variable and their interaction

If x_1 is the exposure of interest, coded as present=1, absent=0, x_2 is a continuous covariate, and their interaction is to be included in the model, the odds of disease if $x_1 = 1$ and $x_2 = x_2$ is $e^{\beta_0+\beta_1+\beta_2x_2+\beta_3x_2}$. The odds of disease when $x_1 = 0$ and $x_2 = x_2$ is $e^{\beta_0+\beta_2x_2}$. The odds ratio is thus $e^{\beta_1+\beta_3x_2}$, ie the odds ratio does not only depend on the coefficient of the exposure of interest.

2.5.6 Summary

In certain cases the coefficients of the variables in the logistic regression

model have direct interpretation as the log odds ratio. This is the case when a dichotomous variable has been coded as 1=present, 0=absent; when a k-category variable has been coded using reference cell coding, and in the case of a continuous variable, if no interactions terms are fitted in the model. Since the coding of categorical variables determines how the coefficients can be interpreted, it must be stated how variables were coded, and it is best to use reference cell coding (Lemeshow and Hosmer 1984).



2.6 The rationale for variable selection

Following the approach outlined in Section 2.3 to Section 2.5, it is fairly straightforward to estimate and interpret the coefficients of a given set of predictor variables. The choice of which variables to include in the set (variable selection, also called model selection) is, however, far from straightforward. In medical applications of logistic regression, information on numerous characteristics of the patient/individual (physical, treatment, exposure to risk factors) may be available. Then the problem of variable selection exists whether one wants to investigate the effect of a certain risk factor, while adjusting for the effect of other factors, or whether one wants to form a prediction equation from a number of potential predictors. In the former case one has to decide which variables to include as confounders, in the latter which variables to select as predictors.

All variables thought to play a role in prediction or confounding should form the pool of variables from which to select a model. As Draper and Smith (1981) point out in the linear regression situation, if it is found that variables which are expected to be important cannot be measured, and that there are no substitute variables which can be used in their place, the model selection has little chance of leading to a useful model, and no model selection should be attempted.

Why do variables have to be selected? As Hosmer and Lemeshow (1989) point out, the fitting of all possible predictors may lead to overfitting and

numerically unstable estimates. By minimising the number of variables in the model one is more likely to obtain a numerically stable and easily generalisable prediction equation than by including all potential predictors in the model. It also makes practical sense, in the case of forming a prediction equation which will be used on future samples, to select an equation containing only important predictors, so that in future information need not be collected on all possible predictors.

On the other hand, Rothman (1986), who is concerned with effect estimation rather than the forming of prediction equations, states that the goal of selecting a parsimonious model with respect to the number of variables selected "is not pertinent to epidemiologic analysis that is focusing on the effect of specific factors". According to him the simplicity of the model is not an important goal in epidemiology: "the primary advantage of employing a multivariate model in an epidemiologic analysis is the ability to control efficiently for a multitude of factors simultaneously". Most texts, however, stress that the accuracy with which parameters are estimated depends on the number of parameters relative to the number of observations. Fitting too many parameters to too few observations leads to overfitting and thus instability. Such a model reflects "more about the particularities of its corresponding data set than about the underlying phenomenon. The hazards of using such models for interpretation or prediction are obvious" (Linhart and Zucchini 1986).

In selecting a subset of variables from a larger set, one thus has to weigh up bias against variance. The smaller the set of variables chosen the more precise (least variable) the estimates but the estimates may be biased. On the other hand, including many variables will reduce the chance of bias but increase the variance. Most variable selection procedures aim for a compromise between variance and bias. Hypothesis testing aims to identify "important" predictors, thus minimising bias as well as variance (since not all variables are fitted). Other approaches aim to minimise a function of bias and variance,

for example the mean square error, prediction error, or error rates. Variable selection procedures will be discussed in the following two chapters.

CHAPTER 3: VARIABLE SELECTION FOR PREDICTION

In this chapter a detailed description of the approaches which have been proposed in the statistical, medical and epidemiological literature for selecting variables in prediction problems will be given. Most of the approaches fall into two distinct categories, namely either hypothesis testing (where variables in a sense have to prove their importance or significance), or the comparison of models with respect to some criterion. Some other methods, which for example place importance on the biological relevance of variables, are also described. Table 3.1 gives a summary of the approaches covered. For each approach the method will be outlined, and criticisms discussed in the literature will be outlined. A summary of the strengths and weaknesses of the procedures, as identified by the student on the basis of the literature review, will be given.

Table 3.1: Selection procedures for prediction

Hypothesis testing
- An initial screening on univariate association
- Stepwise selection procedures

Comparing different models on the basis of a criterion
- Akaike's Information criterion, and related criteria
- Error rates
- Best subset linear regression using Mallows' Cp

Purposeful selection using statistical and biological reasoning

Linear combinations of the independent variables

Bootstrap replications to select predictors

Fit all possible predictors

3.1 Hypothesis testing

3.1.1 An initial screening on univariate association

Some authors suggest that an initial screening of the potential predictors should be done by investigating the univariate associations of each predictor with the outcome, through chi-squared tests in the case of categorical variables and t-tests or univariate logistic models in the case of continuous variables (Hosmer and Lemeshow 1989). In the case of categorical variables a contingency table of the variable against outcome will indicate whether there are any zero cells. A zero cell can be eliminated by collapsing categories in a biologically meaningful way (Hosmer and Lemeshow 1989). Including a variable with a zero cell in the later multiple logistic regression can cause problems.

All variables with $p<0.25$ (suggested by Hosmer and Lemeshow 1989), or $p<0.10$, or $p<0.05$, where the p-value is based on these univariate tests of association, should be considered possible predictors in further selection. Hosmer and Lemeshow (1989) point out that the restriction to variables with $p<0.05$ may exclude variables known to be important. They further stress that variables which are known to be biologically important should be added to those which satisfy the criterion of having $p<0.25$.

The univariate approach does not provide information on subgroups of variables, each of which may be only weakly associated with outcome, but together may form an important set of predictors. Hosmer and Lemeshow (1989) propose that the univariate significance level be increased if it is suspected that there may be such subgroups.

3.1.2 Stepwise selection (forward)

Stepwise selection procedures which select or delete variables from a model are based on algorithms which evaluate the statistical significance of the variables, by means of the likelihood ratio chi-square test.

In forward selection the procedure starts by fitting only the intercept term. Thereafter, for each of the p possible predictors, a univariate logistic regression containing the intercept and that predictor (say x_j) is fitted. The log-likelihood of the intercept model (L_0) is compared with the log-likelihood of each of the univariate models (L_j) by means of the likelihood ratio test statistic:

$$G_j = 2(L_j - L_0)$$

The associated p-value is given by $P(\chi_1^2 > G_j)$ if x_j is continuous, and $P(\chi_{k-1}^2 > G_j)$ if x_j has k categories. The variable identified as the most important at this step is the one with the smallest p-value. This variable (x_e) is thus entered into the model.

In the next step all $p-1$ models containing the intercept, x_e, and one of the x_j ($j \neq e$) variables are fitted. The log-likelihoods of these models are compared with that of the model containing the intercept and x_e. The variable with the smallest p-value at this step is entered, and the algorithm continues.

To use this algorithm a cutoff point for the p-value to enter must be specified. As with the univariate screening outlined in Section 3.1.1, most authors propose a cutoff of greater than the usual 0.05, rather 0.15 or 0.20. The variable with the smallest p-value smaller than the cutoff will thus be entered at a given step. If no variable has a p-value smaller than the cutoff, the selection procedure stops at that step.

Important variables can be forced into the model and selection from the remaining variables proceeds as outlined above, after a model containing the forced variables is fitted at the first step. Variables are thus considered on the basis of the contribution they make to the prediction, in the presence of the forced variables.

3.1.3 Stepwise selection (backwards)

In backward elimination a full model containing all predictors is fitted

and all variables are then evaluated to see which one should first be eliminated from the equation since it makes no or very little contribution to the prediction. At the first step the log likelihood of the model containing all variables (L_f) is compared to that of the p models containing all variables but one (L_{-j}) by the likelihood ratio test statistic

$$G_{-j} = 2(L_f - L_{-j}).$$

The variable to remove from the model is the one which, when removed, yields the largest p-value. At the next step the log likelihood of the model excluding the one removed at the previous step is compared to those of all $p - 1$ models with one of the remaining variables removed. Deletion of variables proceeds as before.

A cutoff value for the p-value to remove must be specified. P-values to leave are generally 0.20 or 0.25. Variables which are known to be important predictors can be forced to remain in the model and only the other possible predictors are considered as candidates for elimination.

Backward elimination is recommended by some authors since starting with all the variables in the model provides some assurance that variables which are important only after adjustment for other variables will not be missed (Hosmer and Lemeshow 1989).

3.1.4 Stepwise selection

In stepwise selection which can start with a full model, with the model containing no predictors, or with a model containing some forced variables, variables which have been eliminated can again be considered for inclusion, and variables already included in the model can be eliminated. It is important that the p-value to leave is defined to be greater than the p-value to enter, otherwise the algorithm could enter and delete the same variable at consecutive steps. Variables can be forced to remain in the model and only the other variables are considered for elimination or inclusion.

3.1.5 Criticisms of stepwise selection procedures

Many authors have criticised these methods: as Draper and Smith (1981) state for linear regression "the screening of variables should never be left to the sole discretion of any statistical procedure". These techniques are criticised because they can lead to statistical models which are biologically implausible. Furthermore, many authors have pointed out that these techniques will not necessarily select the best predictors, and that the ordering of variables obtained is an artefact of the algorithm. If two variables are closely related, only one will be selected as "significant", the other will be "not significant". However, Hosmer and Lemeshow (1989) argue that stepwise selection procedures are useful when little is known about possible predictors and interactions. Stepwise procedures are then useful for screening large numbers of variables.

Gordon (1974) states that stepwise procedures are appropriate if the logistic regression is used "synthetically" to identify high risk individuals, but not if the regression is used "analytically" to try and disentangle the contributions of various variables to the outcome.

Hosmer and Lemeshow (1989) point out that the analyst who fails to scrutinise resulting models, and reports the results as the final, best model, is at fault: "the analyst, not the computer, is ultimately responsible for the review and evaluation of the model." As Gordon (1974) phrases the problem: "the power and elegance of the logistic function make it an attractive and flexible statistical instrument, but in the end, we cannot push a button and hope that everything will come out all right. Because frequently, it will not."

3.1.6 Reporting of stepwise analyses (Hauck and Miike 1991)

Hauck and Miike (1991) have proposed a way of examining and reporting stepwise analyses so that one does not fall in the trap of ignoring correlations between the predictors, and calling the selected variables "the important variables" and the excluded variables "not significant". This is

done by identifying, at each step, close alternatives for entering (in the case of the selection procedure starting with no variables entered): variables that were statistically significant before the step but became non-significant after another variable entered. The authors point out that this criterion for close alternatives must be used with judgement: a change from 0.045 to 0.055 does not change one's perception of that variable, so the variable would not be considered a close alternative. The approach can also be used in backward elimination where the criterion for a transition is reversed: a variable which was non-significant becomes significant.

The presentation is aimed to encourage researchers to think in terms of variables that, together, represent factors associated with outcome (for example education level and income both represent aspects of socio-economic status). An alternative to this procedure would be to ask the researcher to identify variables which are close proxies for one another and then enter only one variable from such a group, or to repeat analyses, first with the one variable and then with the other (Gordon 1974).

3.2 Comparing different models on the basis of a criterion

In the terminology of Linhart and Zucchini (1986) the operating model is the model which we "use to think about the data", the nearest representation of the true situation. There may, however, not be enough information to specify the operating model fully, one may only be able to describe the operating family.

To fit a model one can use a simpler approximating family from which the final model is chosen. The accuracy of any model is measured by a discrepancy, a measure of the lack of fit of the model at hand relative to the operating model. The model which is estimated to minimise the expected discrepancy is the final ("best") model chosen. The overall discrepancy consists of two components: discrepancy due to approximation (bias) and discrepancy due to estimation (variance). The discrepancy due to approximation decreases as

the number of parameters increases and the approximating family approaches the operating family; the discrepancy due to estimation increases as the number of parameters increases. The operating model, being the model with the largest number of parameters, therefore does not necessarily yield the most accurate model.

Depending on the objectives of the analysis, an appropriate discrepancy is defined. A consistent estimator of the expected discrepancy is called a criterion, and is used for model selection. For the selection procedure the criterion can be derived by finite sample methods, asymptotic methods, bootstrap methods, or cross-validatory methods.

3.2.1 Akaike's Information Criterion, and related criteria

A class of model selection criteria is given by

$$C = D + \alpha p$$

where D is the deviance of the model at hand, and p is the number of parameters estimated; α is a constant, or a function of the number of observations (n) (McCullagh and Nelder 1989). [Other authors express such a criterion as maximising log likelihood $-\alpha p/2$ (Atkinson 1980, Stone 1977).] Such criteria take into consideration that the unreserved maximisation of the likelihood provides an unsatisfactory method of choice between models which differ in the number of parameters included (Stone 1977). Atkinson (1980) suggests that α should be varied between 2 and 6, and the effect of these changes on the models chosen should be investigated.

Akaike's information criterion is obtained for $\alpha = 2$, that is

$$AIC = D + 2p$$

or AIC=-log likelihood + p or AIC=(-log likelihood +p)/n. The discrepancy which this criterion estimates is the Kullback-Leibler discrepancy (Linhart and Zucchini 1986 p.18).

Different models can be fitted and compared on their values for this criterion. The model with the smallest value for this criterion is taken as the "best" model, and if there are various models with similar small values for the criterion, these are taken as the best models.

Another criterion based on the log-likelihood is BIC, defined as

$$BIC = D - \text{ (degrees of freedom)}\ln n$$

(Raftery 1986). Raftery (1986) defines BIC as $-2\ln B$ where B=Prob[M_0 is the true model given the data]/Prob[M_1 is the true model given the data], thus the criterion is defined in the context of Bayesian analysis. One should thus specify prior beliefs about the models and their parameters. However, in large sample situations the effect of the prior beliefs would be negligible (Raftery 1986). This criterion is equivalent to the criterion defined by Schwarz as maximising the log-likelihood $-(k\ln n)/2$, where k is the number of parameters and n the number of observations (Schwarz 1978).

3.2.2 Error rates

Ideally the model selected on the basis of a data set (training sample) should be validated on an independent sample (test sample) to see how accurately the model predicts. Few researchers are in the fortunate position to be able to do so. One alternative would be to use part (for example half) of the data set to select a model, and then determine how well this model performs on the remaining observations. However, sample size places constraints on how many variables can be fitted, and the approach of splitting the data set in two is thus only realistic if the whole data set is large. Spiegelhalter (1986) outlines how the explicit aim of predicting the outcome of future observations need to be taken into account to derive useful prediction equations for patient management. He assesses the predictive performance of various models on a test sample by decomposition of a scoring rule. Models selected on the basis of a procedure such as stepwise selection do not necessarily give the best future prediction.

Most researchers use a data set to derive a prediction equation, and then determine the model's prediction accuracy on the same data set, for example by defining observations with a predicted probability of success of more than 0.5 as predicted successes, and observations with a predicted probability of 0.5 and less than 0.5 as predicted failures. The apparent error rate (Efron 1986), calculated in this way, is an underestimate of the true error rate since the model selected was exactly the one which closely predicted the observed data. The apparent error rate can be defined in many ways, for example as the counting error (also called classification error)

$$\frac{1}{n}\sum_{i=1}^{n}(y_i[\hat{\pi}_i \leq 0.5] + (1-y_i)[\hat{\pi}_i > 0.5])$$

(Efron 1986) or

$$\frac{1}{n}\sum_{i=1}^{n}(y_i[\hat{\pi}_i < 0.5] + (1-y_i)[\hat{\pi}_i > 0.5] + 0.5[\hat{\pi}_i = 0.5])$$

(Van Houwelingen and Le Cessie 1990) where [.] represents the indicator function so that for example $[\hat{\pi}_i < 0.5]$ is equal to one if $\hat{\pi}_i < 0.5$ and zero otherwise.

The apparent squared error is given by

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\pi}_i)^2$$

and the apparent deviance by

$$-2\frac{1}{n}\sum_{i=1}^{n}(y_i \ln(\hat{\pi}_i) + (1-y_i)\ln(1-\hat{\pi}_i))$$

which is twice the mean value of minus log-likelihood described by Van Houwelingen and Le Cessie (1990).

What one would wish to estimate is the actual error rate, the error rate obtained by averaging over the distribution of future observations. Resampling techniques such as the bootstrap and cross-validation (Efron and

Tibshirani 1993) can be used to estimate the actual error rates of various models using only the original data set.

In cross-validation, each observation is removed from the data set, one at a time, and the model is fitted to the remaining $(n-1)$ observations. The fitted model is then used to predict the outcome of the eliminated observation (for each of the n observations). An observation is thus excluded from the construction of the model for its own prediction. Cross-validation error rates of various models can be calculated in this way. If $\tilde{\pi}_i$ denotes the predicted value for observation i, based on a model derived from the $n-1$ observations excluding observation i, the cross-validation estimate of the actual counting error is

$$\frac{1}{n}\sum_{i=1}^{n}(y_i[\tilde{\pi}_i \leq 0.5] + (1-y_i)[\tilde{\pi}_i > 0.5])$$

or

$$\frac{1}{n}\sum_{i=1}^{n}(y_i[\tilde{\pi}_i < 0.5] + (1-y_i)[\tilde{\pi}_i > 0.5] + 0.5[\tilde{\pi}_i = 0.5])$$

An alternative method of obtaining estimates of the actual error rates is the bootstrap. Many random samples (say 1000) of size n are drawn with replacement from the data set of size n. If B denotes the number of bootstrap samples taken, the bootstrap estimate of the actual counting error is

$$\frac{1}{Bn}\sum_{j=1}^{B}\sum_{i=1}^{n}(y_i[\hat{\pi}_{ij} \leq 0.5] + (1-y_i)[\hat{\pi}_{ij} > 0.5])$$

or

$$\frac{1}{Bn}\sum_{j=1}^{B}\sum_{i=1}^{n}(y_i[\hat{\pi}_{ij} < 0.5] + (1-y_i)[\hat{\pi}_{ij} > 0.5] + 0.5[\hat{\pi}_{ij} = 0.5])$$

Efron (1983) describes various other ways of estimating the actual error rate, for example the double bootstrap.

Van Houwelingen and Le Cessie (1990) provide an approximation for the expected optimism of the mean value of minus log-likelihood error rate (half the apparent deviance), namely p/n. The expected actual mean value

of minus log-likelihood error rate is thus

$$-\frac{1}{n}\ln L + \frac{p}{n}$$

which is essentially Akaike's information criterion (Linhart and Zucchini 1986). Efron (1986) derives an approximation for the expected apparent error rate, based on counting error.

Titterington et al (1981) point out that error rates are insensitive since no account is taken of the relative seriousness of different errors. Van Houwelingen and Le Cessie (1990) point out that error rates are generally not used to compare different models, but to describe the success of one prediction rule.

3.2.3 Best subsets linear regression using Mallows' C_p (Hosmer, Jovanovic and Lemeshow 1989)

A discrepancy often used for selection in linear regression is the average mean squared error of prediction (Linhart and Zucchini 1986 p116). A criterion which estimates this discrepancy is Mallows' C_p

$$C_p = RSS_p/s^2 - (n - 2p)$$

where RSS_p is the residual sum of squares from a model containing p parameters and s^2, the residual mean square from the model containing all the predictors, is presumed to be a reliable unbiased estimate of the error variance σ^2 (Draper and Smith 1981). If a model containing p parameters does not suffer from lack of fit C_p will be close to p.

Hosmer, Jovanovic and Lemeshow (1989) have outlined how best subsets linear regression programmes using C_q (as they prefer to call C_p) as criterion for selection can be used to do best subsets selection of logistic regression models. After fitting the logistic regression model containing all possible predictors, a modified dependent variable and case weights are obtained using the predicted values of the full logistic model. Best subsets linear regression

is then performed on this modified dependent variable with case weights: as outlined in Estimation (Section 2.3) $\hat{\beta}$ can be obtained by using a method of iteratively reweighted least squares, with modified dependent variable

$$z_i = ln\frac{\hat{\pi}_i}{1-\hat{\pi}_i} + \frac{y_i - \hat{\pi}_i}{\hat{\pi}_i(1-\hat{\pi}_i)} \qquad i = 1, \ldots, n$$

and weight function

$$w_i = \hat{\pi}_i(1-\hat{\pi}_i).$$

Then $\hat{\beta} = (X'WX)^{-1}X'Wz$. If z_i is the dependent variable, x_i are the predictors, and w_i the case weights, a linear regression programme will give estimated coefficients identical to $\hat{\beta}$. This relationship provides the basis for model selection using linear regression techniques.

If the fitted value of the i-th case from the linear regression with all p predictors is $z_i(p) = x_i'\hat{\beta}$, the residual sum of squares from the weighted linear regression is

$$\begin{aligned} SSE(p) &= [z - \hat{z}(p)]'W[z - \hat{z}(p)] \\ &= \sum_{i=1}^{n} w_i[z_i - \hat{z}_i(p)]^2 \\ &= \sum_{i=1}^{n} \hat{\pi}_i(1-\hat{\pi}_i)\frac{(y_i - \hat{\pi}_i)^2}{\hat{\pi}_i^2(1-\hat{\pi}_i)^2} \\ &= \sum_{i=1}^{n} \frac{(y_i - \hat{\pi}_i)^2}{\hat{\pi}_i(1-\hat{\pi}_i)} \end{aligned}$$

This is the Pearson chi-square goodness of fit statistic, X^2, for the fitted logistic regression model. The estimated covariance matrix for the estimated coefficients computed by the linear regression is

$$\frac{X^2}{(n-p-1)}(X'WX)^{-1}.$$

The estimated standard errors from the logistic regression model are thus the estimated standard errors from the linear regression model, divided by $[X^2/(n-p-1)]^{\frac{1}{2}}$, the square root of the mean square error. To fit a subset q of

the predictors in a linear regression model, X is partitioned as $X = (X_1, X_2)$ with X_1 an $n \times (q+1)$ design matrix and X_2 an $n \times (p-q)$ matrix. β' is similarly partitioned as $(\beta_1'\beta_2')$. z and w are computed using $\hat{\pi}$ from the full logistic model. $X'WX = I$ is partitioned as

$$\begin{pmatrix} I_{11} & I_{12} \\ I_{21} & I_{22} \end{pmatrix}$$

with

$$\begin{aligned} I_{11} &= (X_1'WX_1) \\ I_{12} &= (X_1'WX_2) = I_{21}' \\ I_{22} &= (X_2'WX_2) \end{aligned}$$

Fitting the linear regression with z as dependent variable, X_1 as predictors and W the weight matrix gives

$$\tilde{\beta}_1 = (X_1'WX_1)^{-1}X_1'Wz$$

which can be shown to equal

$$\tilde{\beta}_1 = \hat{\beta}_1 + I_{11}^{-1}I_{12}\hat{\beta}_2$$

If the vector of fitted values from the linear regression model with q variables is $\tilde{z}(q) = X_1\tilde{\beta}_1$, the residual sum of squares for the model containing the q variables is

$$\begin{aligned} SSE(q) &= [z - \tilde{z}(q)]'W[z - \tilde{z}(q)] \\ &= z'Wz - \tilde{\beta}_1'(X'WX)\hat{\beta}_1 \\ &= X^2 + \hat{\beta}'(X'WX)\hat{\beta} - \tilde{\beta}'(X_1'WX_1)\tilde{\beta}_1 \\ &= X^2 + \tilde{\beta}_2'(I_{22} - I_{21}I_{11}^{-1}I_{12})\hat{\beta}_2 \end{aligned}$$

The increase in the residual sum of squares when excluding $p-q$ variables is given by $\lambda^* = \hat{\beta}_2'(I_{22} - I_{21}I_{11}^{-1}I_{12})\hat{\beta}_2$. Since $(I_{22} - I_{21}I_{11}^{-1}I_{12})$ is an estimate

of the inverse of the covariance matrix of $\hat{\beta}_2$, λ^* is the unconditional Wald test statistic for testing the hypothesis $H_0 : \beta_2 = 0$.

To compare the fit of the different models C_q is proposed. In linear regression

$$C_q = \frac{SSE(q)}{SSE(p)/(n-p-1)} + 2(q+1) - n$$

Thus in the setting outlined above

$$C_q = \frac{X^2 + \lambda^*}{X^2/(n-p-1)} + 2(q+1) - n$$

Under the assumption that the model is the correct one, X^2 has approximate expected value $(n-p-1)$ and λ^* has approximate expected value $(p-q)$. Thus

$$\begin{aligned} E[C_q] &= (n-p-1) + (p-q) + 2q + 2 - n \\ &= q + 1 \end{aligned}$$

If the subset of variables excludes important variables, λ^* will follow a non-central χ^2 and C_q would be larger than $q+1$. Models with C_q near $q+1$ are candidates for a best model. The best subsets linear regression programme will select as best the subset with smallest value of C_q.

McCullagh and Nelder (1989) propose a correction factor to adjust the expectation of X^2 but Hosmer, Jovanovic, and Lemeshow (1989) recommend the use of C_q as outlined above, until the correction factor has been investigated further.

Based on the work of Hauck and Donner (1977) who examined the inferential adequacy of λ^*, and found that λ^* may fail to reject the hypothesis that all $p-q$ coefficients in β_2 are zero, especially when n is small, C_q can be expected to be small for subsets of variables whose coefficients are not all equal to zero. A larger number of subsets than those indicated by the conservative values of C_q should therefore be considered.

Because of the approximate nature of the estimated coefficients from the best subsets linear regression Hosmer, Jovanovic and Lemeshow (1989)

propose that all selected models be refit using a logistic regression programme to obtain the correct estimates. It is also stressed that the best subsets linear regression should be used to select a core of important covariates from the full set and that these models should be critically evaluated. For example, a biologically important variable may be forced into the model irrespective of the results of the subset selection procedure.

3.3 Purposeful selection using statistical and biological reasoning

This selection method attempts to address the criticisms levelled at stepwise procedures, namely that the computer is doing the selection, rather than the analyst (Gordon 1974, Hosmer and Lemeshow 1989). Various models are fitted and have to be evaluated statistically as well as clinically.

Hosmer and Lemeshow (1989) propose that model selection should start with the full model (or, if there are too many variables relative to the number of observations, by a model containing all variables identified as statistically relevant on the univariate analysis, or biologically important, or those identified by a best subsets selection). After fitting the full model, the Wald statistic for each variable should be examined: the Wald statistic compares the estimate of the coefficient to its estimated standard error. Under the hypothesis that $\beta_j = 0$ the statistic asymptotically follows a standard normal distribution. The estimated coefficient from the full model should also be compared to that of the univariate model. Variables that do not contribute on the basis of these criteria should be eliminated. The new model is then compared with the old one by the likelihood ratio test. The coefficients of the variables retained in the new model are compared with their coefficients in the full model: if there are marked changes it would indicate that the deleted variables provided needed adjustments. The process of deleting, fitting and verifying continues until only biologically or statistically important variables are retained in the model. At this stage the appropriateness of the assumption that all continuous variables are linear in the logit should be examined

(see Section 3.6.2).

The need for including interaction terms is assessed hereafter. It is suggested that only interactions which have a biologic foundation should be investigated. Furthermore, if a variable is not found to contribute in the full model, but it is expected that the variable interacts with some others (age is an example) the variable is retained so as to be able to investigate the presence of interactions. The likelihood ratio test is used to determine whether an interaction term is at least moderately significant: interaction terms should be biologically plausible and statistically significant. Having assessed the fit and adequacy of the model, the model is interpreted in terms of odds ratios of the predictors.

Collett (1991) and Cox and Snell (1989) propose a similar variable selection procedure. It is stressed that there are probably different models which fit equally well. These models should, if possible, all be reported and any choice between alternatives should be made on clinical grounds.

3.4 Linear combinations of the independent variables

An approach which potentially uses information of all variables but reduces the number of variables entered into the model is to group variables into linear combinations and then use these linear combinations as possible predictors. As has been stated in a different regression situation "it is better to use a little bit of all the variables than all of some variables and none of the remaining ones" (Marquardt and Snee 1975). Variable clustering can be used to create groups of variables such that the variables within a group are highly correlated. Variables within such a group would usually represent the same clinical phenomenon but clinicians can also assist in forming groups of variables that belong together. In practice variables with small weights in the linear combination are often excluded.

3.5 Other methods

3.5.1 Bootstrap replications to select predictors (Efron and Gong 1983)

Efron and Gong (1983) outline how the bootstrap can be used to determine which variables from a larger set are significant predictors. For each bootstrap sample (in their case 500 samples) drawn from the training set (in their case n=155), a prediction rule consisting of three steps was followed (at each step only observations with complete data for the variables being considered were included in the analysis):

1. For each of the possible predictors (in their case 19) a univariate logistic regression model was fitted. Predictors for which

$$H_0 : \beta_j = 0$$

was significant at the 0.05 level, were selected for consideration in the next step.

2. The predictors found to be significant in step 1 were entered into a forward selection programme starting with only the constant term. Entry level was defined as 0.10. This step stopped when no further variables achieved the significance level.

3. The variables selected in step 2 were entered into a forward selection programme with entry level defined as 0.05. Variables which were selected in this way constituted the final model.

Efron and Gong found that, of the 4 variables chosen as important when the data set was first analysed (using the above prediction rule on the data set), not one was selected as important in more than 60% of the bootstrap samples: one was selected in 37% of the replications, one in 48%, one in 35% and one in 59%. Such a result clearly indicates that there is not one best model to be selected.

As Van Houwelingen and Le Cessie (1990) point out, such a method, or a similar method using cross-validation, needs very strict rules for model build-

ing, rules which are hard to provide since much of model building proceeds by trial and error.

3.5.2 Fit all possible predictors

A method which is not proposed in the literature but is used frequently in the medical literature consists of fitting all possible predictors. The importance of variables is evaluated by means of p-values or confidence intervals for odds ratios in the full model. The motivation given at the beginning of this chapter for variable selection in general, can be used as criticism of this approach.

3.6 Other issues

Other guidelines given by Hosmer and Lemeshow (1989) deal with numerical problems, such as zero frequency in a contingency table, predictors discriminating perfectly (ie there is no overlap between the two outcome groups in the distribution of the predictors, possibly only due to a numerical coincidence), and colinearity. These problems always lead to extremely large estimated standard errors and sometimes to large estimated coefficients.

3.6.1 Sample size

It is often mentioned that the number of variables that one can consider depends on the sample size one is dealing with. So, for example, in a discussion on the merit of starting variable selection with a multivariate model containing all possible variables, Hosmer and Lemeshow (1989) point out that the appropriateness of this approach "depends on the overall sample size and the number in each outcome group relative to the total number of candidate variables". No clearcut guidance is given, however, and various rules of thumb exist, for example the sample size should be at least 4 times, or at least 10 times the number of variables, or there should be 10 of the rarer outcomes for each variable fitted.

3.6.2 The scale of a continuous predictor

When a continuous predictor is considered for inclusion in a model, it has to be determined whether the variable is in fact linear in the logit. Hosmer and Lemeshow (1989) propose that the untransformed variable be used to determine whether the variable is important. Thereafter the need for transformation is considered. They suggest categorising the continuous variable (for example into quartiles) and creating dummy variables (in the case of quartiles creating 3 dummy variables) using the lowest group as the reference. The estimated coefficients of the design variables are plotted against the midpoint of the quartile and from this plot the most appropriate choice of transformation is selected.

An alternative is to include the continuous variable in the model, as well as a term consisting of the product of the variable and its logarithm. If the latter term makes a significant contribution the variable is nonlinear in the logit.

Collett (1991) outlines a variety of residual plots which can be used to investigate the need for transformations, and Royston (1992) describes the use of cusum plots.

Greenland (1983) points out that it is common practice to use untransformed continuous variables but that this has no justification other than convenience. To circumvent the technical problems of continuous variables many authors (for example Rothman 1986) recommend categorisation of continuous variables.

If a continuous variable is categorised one has to decide where the category boundaries should be drawn. This decision should be meaningful clinically, may depend on the distribution of the data, but should not be made in such a way as to influence the results in a certain direction (Rothman 1986).

3.7 Summary of selection procedures

The selection approaches outlined above vary considerably as far as com-

putational and intellectual complexity are concerned. Some are computationally straightforward with the analyst in essence pushing a button and a standard computer package such as SAS or BMDP doing the work. On the other hand, some entail extensive programming (for example to be able to do resampling from the data set) or a lot of time to fit and evaluate various models on the basis of some criterion. Table 3.2 gives a summary of the complexity of the different procedures, whether criticisms have been levelled against each procedure and whether the procedures are used in the medical literature.

Table 3.2: Summary of selection procedures for prediction

Procedure	Computationally	Intellectually	Criticisms	Used
hypothesis testing	easy	easy	many	often
comparing criteria	time-consuming	easy	some	seldom
purposeful selection	easy	difficult	none	seldom
linear combinations	?	?	some	seldom
bootstrap replications	difficult	easy	some	never
fit all predictors	easy	easy	many	often

As the aim of the prediction analysis is mostly to determine which are important predictors, and not to come up with a synthetic prediction equation (Gordon 1974) the approaches should be evaluated in terms of their ability to identify important predictors. As a point of departure it has to be accepted that there is often no one best model, but that there may be various useful sets of variables (Cox and Snell 1989). Selection procedures which take this into account are thus more desirable. Therefore stepwise procedures which select one final model would seem to be inappropriate. It is, however, not necessarily the procedure which is inappropriate, but rather the way in which the procedure is used. If the guidelines of Hauck and Mike (1991) regarding the identification of close alternatives for variables are

followed, stepwise procedures can offer valuable insights. Approaches which compare various models on the basis of a criterion seem more desirable since various models with similarly small values of, for example, Akaike's information criterion can be identified. If such methods are used inappropriately, for example to pick the one model with the smallest value (ignoring others which have values which are not much larger), they can, however, also lead to incorrect answers. Most importantly, researchers have to be educated that there need not be one best model, and that the analyst is doing a proper job when various alternatives are identified. Often researchers want one final (simplistic) answer, and are rather taken aback when the analyst is not prepared to say what the final, best answer, is.

If synthetic prediction is the aim linear combinations of predictors are useful. If, however, a set of important predictors are to be identified other procedures are clearly more useful.

The intellectual contribution in model building often has to be provided by the researcher, by specifying which are biologically important variables, and which interactions are expected. If the researcher is not capable of such input the analyst has difficulty in building a biologically plausible and useful model.

Given that there is such a wide range of possible approaches to predictor selection it is interesting to note that an approach used very often is the one which consists of fitting all predictors in one model and identifying variables that have significant p-values or significant confidence intervals for the odds ratio in the full model as significant. Computationally and intellectually this is of course the easiest approach. Of the 34 articles and briefs using logistic regression for prediction published in the American Journal of Public Health in 1992, 22 used this approach. Problems of the imprecision of such estimates are not discussed.

More specific comments on the various approaches will be given in Chapter 6 where the approaches will be tried out on various data sets.

3.8 Concluding remarks

The steps of model development outlined by Draper and Smith (1981) for linear regression could be applied similarly for logistic regression:

-collect data, check quality of data, plot, try models

-consult experts in the subject field for criticism

-validate the model

The importance of the data checking phase cannot be stressed too much. The model-building process proposed by Hosmer and Lemeshow (1989) starts with univariate analyses, which will enable the researcher to easily detect errors or extremely small frequencies, which, if not corrected (in the case of errors) or not excluded from the analysis (in the case of extremely small cells) could lead to strange results in the multiple logistic regression. Rothman (1986) criticises multiple regression techniques because of the barrier they place between the researcher and the data, hindering the researcher from getting to know the data. This is of course only true if the researcher launches into a multiple regression analysis before first investigating the data univariately, an approach which should not be used.

CHAPTER 4: VARIABLE SELECTION FOR ESTIMATION

In this chapter the concepts of confounding and effect modification will be discussed. Thereafter, the approaches which have been proposed in the statistical, medical and epidemiological literature for selecting variables in estimation problems in logistic regression will be described. Table 4.1 gives a summary of the approaches covered. For each approach the method will be described, and criticisms discussed in the literature will be outlined. A summary of the strengths and weaknesses of the procedures, as identified by the student on the basis of the literature review, will be given.

Table 4.1: Selection procedures for estimation

Significance testing
Fit all known confounders
Combine prior beliefs and data estimates
Select on basis of change in the estimate of interest
Composites of possible confounders
Minimise the error of estimation
Partial Gauss discrepancy

4.1 Confounding

4.1.1 The definition of confounding

Confounders are factors which are associated with both the exposure of interest and, independent of this association, with the outcome (disease) under consideration (Rothman 1975). In case-control studies confounders are factors which are associated with exposure in the control group, since they are taken as a surrogate for the population from which cases and controls are drawn (Hauck et al 1991). The presence of such factors distorts the association which exists between the exposure and disease. Some authors

refer to these as classical confounders (Hauck et al 1991). Example 4.1 shows the presence of a classical confounder.

Example 4.1

In 1990 a study was conducted on adult inhabitants of QwaQwa in the Orange Free State (Mollentze et al 1994). Information was gathered on risk factors for cardiovascular disease. As one of the analyses the researcher wished to investigate the association between obesity (body mass index>25) and hypertension in adults aged 45 years and older. Sex is a possible confounder since women are more likely to be obese, and women are more likely to be hypertensive. Sex is thus associated with the exposure and the disease, but is clearly not a consequence of the exposure. If the odds ratio of hypertension (obese relative to non-obese) is calculated from the following 2x2 table ignoring sex, or from fitting a logistic model containing only obesity, the crude odds ratio is 2.2.

	Hypertensive	Not hypertensive	
Obese	151	108	259
Not obese	84	132	216
	235	240	475

Stratifying by sex the results are as follows, and the Mantel-Haenszel adjusted odds ratio is 1.8 (which is also obtained from fitting a logistic model containing obesity and sex).

	Males			Females		
	hyp	not hyp		hyp	not hyp	
obese	20	24	44	131	84	215
not obese	36	73	109	48	59	107
	56	97	153	179	143	322

For males the odds ratio is 1.7 and for females 1.9, thus both stratum-specific odds ratios are smaller than the crude odds ratio, indicating that the

omission of the confounder sex introduced a slightly stronger relationship than does exist.

It can also be shown that there might be confounders whose omission may mask a real association. So, for example, sex can be a confounder in the association between smoking and hypertension. In this case women have a low prevalence of smoking but a high risk of disease whereas men have a lower risk of disease but a higher prevalence of smoking. The crude estimate of the odds ratio associated with smoking could therefore be much lower than the unconfounded odds ratio.

4.1.2 Controversies regarding the definition of confounding

There is some controversy as to whether an extraneous factor, to be a confounder, should be associated with the disease and exposure in the population (thus based on prior knowledge) or whether the associations should exist in the given data set (thus based on the data). Collett (1991), for example, states that it will only become apparent whether a factor is a confounder once the data has been analysed. On the other hand, Breslow and Day (1980) state that the significance of the association of the possible confounder with the outcome in a given data set is irrelevant: prior knowledge should determine whether a factor is a confounder. Other authors (Day, Byar and Green 1980; Miettinen and Cook 1981) propose that if a factor is known to be associated with the disease but is believed to be not related to the exposure in the population, the factor should be adjusted for since accidents of sampling or design could have created an exposure-confounder association. However, if the factor is known to be related to exposure but not disease in the general population, a confounder-disease association may exist by chance in a given data set, and should not be adjusted for.

A factor is taken to be a confounder if stratification by the confounder alters the association between the exposure and disease (Breslow and Day 1980). However, a factor can do this without being associated with the

exposure. Example 4.2, taken from Miettinen and Cook (1981), illustrates that, although the exposed and unexposed did not differ with regard to their sex distribution, the odds ratios stratified by sex led to an odds ratio markedly different from the crude odds ratio. Miettinen and Cook (1981) view this not as confounding but as modification, and sees it as an illustration of the peculiarities and subtleties of the odds ratio (stratification would not have altered the estimate of the risk difference, for example). Boivin and Wacholder (1985) feel that there is no conflict if the factor is a confounder with respect to the odds ratio but not the risk difference: one should state beforehand which effect measure is going to be used and then report that confounding was or was not found with respect to this effect measure. Hauck et al (1991) call such confounders mavericks: they do not fit the classical definition of a confounder, but they are operationally confounders in that their inclusion changes the estimate of the odds ratio. Their omission tends to bias the odds ratio towards no effect (Hauck et al 1991).

Example 4.2

The hypothetical data of a followup study with identical sex distributions among the exposed and unexposed are as follows (Miettinen and Cook 1981):

	Males			Females		
	Disease			Disease		
	+	-		+	-	
exposed	99	1	100	5	95	100
unexposed	95	5	100	1	99	100
	194	6	200	6	194	200

The stratum-specific odds ratios are both 5.2. The unstratified data are as follows:

	Diseased	Not diseased	
exposed	104	96	200
unexposed	96	194	200
	200	200	400

The crude odds ratio is 1.2, despite the fact that exposure was not associated with sex.

4.2 Effect modification

4.2.1 The definition of effect modification

If the degree of association between exposure and disease differs for different levels of an extraneous variable, the extraneous variable is called an effect modifier. There is thus an interaction between the variable and the exposure (see Example 4.3). Breslow and Day (1980) say that the medical importance of studying such interactions is to see whether the definition of high risk groups has to be modified, and to gain insights into disease mechanisms. Whereas confounding is a bias which one wishes to remove, one wishes to model and understand effect modification.

Example 4.3

In the survey mentioned in Example 4.1, the researcher was also interested in examining the relationship between insulin and triglycerides. Values of triglycerides are categorised as high if they fall above 2.3. To determine what constitutes high levels of insulin, an approach is to determine the 75th percentile of a subgroup of study participants who are normotensive, not diabetic and not obese (Modan et al 1985). Survey participants who have insulin values higher than the 75th percentile of this reference group are categorised as having high insulin levels. If we consider sex a possible effect

modifier, namely that the effect of insulin on triglycerides differ by sex, and analyse the data of the participants aged 45 and older, stratified by sex, sex is clearly an effect modifier. In females the odds ratio is 1.6, in males 5.2.

	Males Triglycerides			Females Triglycerides		
	high	normal		high	normal	
insulin high	16	113	129	9	33	42
insulin normal	14	160	174	5	96	101
	30	273	303	14	129	143

4.2.2 Controversies regarding the definition of effect modification

Most authors either explicitly or implicitly state that an effect modifier is a confounding variable, in that, having decided which variables are possible confounders, some of them may later be termed effect modifiers since the association between exposure and disease varies for different levels of the confounders (Collett 1991; Hosmer and Lemeshow 1989; Breslow and Day 1980). Miettinen (1974), however, states that an effect modifier need not be a confounder.

4.3 Confounder selection

The controversy regarding the definition of a confounder has been raging in the pages of journals such as the American Journal of Epidemiology for years (for example Miettinen and Cook 1981, Boivin and Wacholder 1985, Grayson 1987). The confusion that reigns in determining what is a confounder is also apparent in the different approaches advocated for selection of confounders to adjust for in a multiple logistic regression. The aim of this part of the thesis is to summarise what is being done regarding confounder (whether classical or operational) selection and give some recommendations.

4.3.1 Introduction

The choice of which variables should be considered possible confounders is inherently Bayesian (Greenland and Neutra 1980, Robins and Greenland

1986): prior knowledge of relationships between exposure and outcome must be used to decide which variables are logical choices for confounders. This is a clinical judgement, not a statistical choice. Since clinical judgement may vary from clinician to clinician it is recommended that it should be stated explicitly which clinical judgements were made in the selection of confounders, possibly through path diagrams which show which relationships are considered causal influences, causally induced correlations and chance correlations (Greenland and Neutra 1980). Variables which are found to be associated with the outcome in the given data set, but are strongly believed to be not related to the outcome, and are not proxies for causally important variables should not be considered possible confounders (Greenland and Neutra 1980; Day, Byar and Green 1980).

Day, Byar and Green (1980) show how adjusting for chance confounders can increase the variability of the estimate of the effect measure of interest. They also point out that bias can occur if one selects from the chance confounders the ones which lead to the largest decrease in the effect measure of interest. Apparent inconsistencies between the findings of studies investigating a given exposure/outcome association may be explained by adjustments having been made for chance confounding.

Schlesselman (1978) proposes a method, using other data sources, to assess the effect of a potential confounder which was not measured in a study. Bross (1966) proposes the Size Rule for similar situations: "strong relationships cannot be explained away by invoking extraneous variables". This type of confounder selection, which requires extensive knowledge about relationships between the confounder, exposure and outcome, which cannot be ascertained from the given data set since the variables were not measured in the study, will not be covered here.

From the pool of possible confounders chosen on clinical grounds, a subset of confounders to be used in the statistical analysis to obtain the most accurate (unconfounded) estimate of the effect measure of interest can

be chosen in a variety of ways.

4.3.2 Significance testing

Significance testing to determine which factors to consider as confounders in the final analysis is used in the following ways:

1) to determine whether the factor is a classical confounder, ie whether the factor is statistically significantly associated with both the exposure and the disease. If the factor is found not to be significantly associated with the exposure or disease, the factor is not considered a confounder;

2) to determine whether the factor is statistically significantly associated with the outcome. If the factor is not significantly associated with the outcome, the factor is not considered a confounder;

3) in stepwise selection procedures to determine which factors are independently of one another associated with the outcome;

4) in collapsibility tests which determine whether the crude odds ratio and the adjusted odds ratios differ statistically significantly (Mickey and Greenland 1989).

Many authors strongly condemn the use of hypothesis tests and p-values for determining whether a variable should be considered a confounder. Breslow and Day (1980) show that a factor which in a given data set is found not to be statistically significantly associated with the exposure or disease can still be considered a confounder when one compares the adjusted odds ratio with the crude odds ratio. Dales and Ury (1978) point out that, when considering the confounding potential of a factor, one wishes to know the likelihood that the associations between the confounder and the exposure, and the confounder and the outcome, are such that the disease-exposure relationship has been appreciably distorted. Sample size, the variability of the observations and the confounder prevalence have pronounced effects on significance testing, whereas they should not have much of an impact on the confounding potential of the factor (Dales and Ury 1978; Rothman 1986).

In large studies many associations would be found "statistically significant", whether the resulting confounding was large or small. In small studies very few significant associations would be found, leading one erroneously to think that there is no confounding present. In significance testing one wishes to prove that the associations do exist. In confounder selection, one should be more interested in proving that the associations do not exist. It is therefore suggested that if significance testing is used, the critical level of 0.25 or even 0.50 is used, instead of the usual 0.05 or 0.01.

Fleiss (1986) counters by stating that significance tests (with $p<0.01$ for the confounder-disease association) to determine which of a large set of possible confounders should in fact be considered confounders in the final analysis provide the explicit and prespecified rules that are required of a reproducible decision-making process. Poole (1987) responds to this by saying it is exactly the hazards of rituals like significance testing that researchers must safeguard themselves against. Significance testing belongs in the realm of decision-making, not in science where one seeks to understand.

Mickey and Greenland (1989), on the basis of a study using Monte Carlo simulation of several confounder selection criteria, conclude that any preliminary testing should only be used in cases where there is little prior information about the confounder effects. They agree that significance levels have to be raised to 0.20 or higher.

Greenland (1989) stresses that if stepwise procedures are used to select independent confounders, the exposure of interest must be forced into the model.

From the definition of what constitutes a confounder it is clear that many authors feel that associations between the confounder and outcome in the data set should play no role in deciding whether a variable should be adjusted for.

An alternative to the above described significance testing which generally are tests of no association, would be to test hypotheses of clinical

relevance (Schall and Luus 1992). To assess the confounder-outcome association, for example, a prespecified value is taken as reflecting a clinically relevant association. This would certainly weed out many nonconfounders identified as confounders on the basis of significance testing in large data sets. In smaller data sets problems with power remain. However, confidence intervals for the association may be useful. In general, epidemiologists are turning away from p-values and towards confidence intervals, since the latter allow one to interpret the magnitude of associations/differences. Confidence intervals for measures of association between possible confounders and exposure/outcome can be clinically interpreted. Whether clinicians would be able to specify what constitutes a clinically relevant association indicating confounding, remains to be seen.

As described above one should decide that a potential confounder is in fact not a confounder only if one can reject the alternative hypothesis that the factor is an important confounder. Greenland (1989) proposes that one should consider equivalence testing if one wishes to perform a statistical test. Equivalence has to be defined as an odds ratio, for example, falling in a given range (of little clinical importance). Hauck and Anderson (1986) propose an equivalence curve for situations where the range of values which are clinically not important is not clearcut.

4.3.3 Fit all known confounders

Instead of selecting a subgroup of the pool of possible confounders, many researchers prefer to fit all possible confounders in the final model. In this way any bias is said to be avoided.

Breslow and Day (1980) state that the significance (statistical or clinical) of the association of the potential confounder with the outcome in the given data set is of no relevance. They propose that a variable which is known to be related with disease (and is not a subsidiary to a possible exposure/disease association) should be treated as a confounder. Therefore variables such as

age and sex should be considered confounders in most applications.

Miettinen and Cook (1981) stress that deciding whether a variable is a confounder involves a priori knowledge and cannot be done by a mechanistic approach which is based only on the relationships in the given data set.

Rothman (1986) states that "there is no compelling reason to reduce the model to a small set of terms" if estimation is the aim. He does concede that the number of terms should not be more than 20% or 30% of the number of observations, but feels that concerns about confounding should dominate one's thinking, rather than simplicity of the model.

Other authors note, however, that the inclusion of all possible confounders may lead to wrong conclusions (Starr, Dalcorso and Levine 1986). The precision of the estimate in a small data set where many possible confounders are fitted, may be questionable. One may also, for many of the possible confounders, have insufficient information to decide a priori whether they are confounders or not, and some selection will have to be done to avoid overadjustment (Day, Byar and Green 1980).

4.3.4 Combine prior beliefs and data estimates

In this approach factors that the researcher strongly believes to be confounders are forced in the model, whereas other possible confounders are screened (in the case of Starr, Dalcorso and Levine (1986) by backward elimination with $p<0.05$) or prior beliefs are adjusted by taking the data into account.

Robins and Greenland (1986) state that modelling strategies should be attempts to approximate Bayesian analysis: one should start with prior beliefs in terms of the effect of confounders, and update these by evidence from the data, since there is rarely enough information in nonexperimental data to allow one to construct an acceptably accurate estimate of the effect of the exposure from the data alone. If the model estimates of parameters are in serious conflict with strong prior beliefs one should not accept the model,

even if it fits the data well. Robins and Greenland (1986) state that forcing certain confounders to be in the model does not go far enough: if one has strong prior beliefs regarding the size of a given coefficient that coefficient should be forced in the model, or a weighted average of the prior belief and the data information should be used. Rothman (1986) states that the selection of factors to enter into a model and the mathematical formulation of the model are not statistical issues but should rely on a biologic understanding of the disease process and the relationships between the factors under study.

If one's prior beliefs are not strong and they are contradicted by the data, one should be willing to give up the prior beliefs. Whether one has enough information to form strong prior beliefs regarding all possible confounders, is questionable (Day, Byar and Green 1980).

4.3.5 Select on basis of change in the estimate of interest

The change-in-estimate criterion (absolute or relative) (Miettinen and Cook 1981) is said, by some authors, to address the essence of confounding: if the inclusion of a variable changes the effect estimate of interest, that variable should be considered a confounder. One thus has to assess the degree of discrepancy between the crude and unconfounded estimates. If the crude and unconfounded estimates are exactly the same, there is no need for adjusting for a confounder. How large a discrepancy has to be to indicate that confounding is present, is problematic. Rothman (1986) states that the comparison between the crude and unconfounded estimates "clearly and unambiguously reveals the magnitude of the confounding, which the investigator can then take into account in further analyses or reporting of results" but gives no further guidance. Some other authors have been more explicit about what is taken to be a sizeable change: Mickey and Greenland (1989), for example, propose a change of 10% as important, acknowledging that this choice is somewhat arbitrary but countering that it is no more arbitrary than the choice of significance levels for the statistical tests. Hosmer

and Lemeshow (1989) propose that a biologically important change in the estimate of the coefficient should be taken as indicating the presence of a confounder.

Greenland (1989) points out that the change-in-estimate criterion takes no account of the random variability of the two estimates being compared. Selection should therefore rather be based on change-in-confidence-intervals of the estimate.

The change-in-estimate criterion can lead to false conclusions (Miettinen and Cook 1981). Clearly, if one investigates all extraneous variables to see whether their inclusion has an effect on the estimate of the effect measure of interest, a variable may, by chance, have an effect, although there is no theoretical basis for the association. Miettinen and Cook (1981) thus criticise this criterion, stating that it is totally dependent on the data at hand. Restricting the pool of possible confounders to only those which have some theoretical basis, would remove this problem.

4.3.6 Composites of possible confounders

Miettinen (1976) proposes that a multivariate confounder score should be constructed for each individual. Discriminant analysis or logistic regression can be used to form a function which separates the cases and the controls, or the exposed from the non-exposed. The subjects are then divided into a few strata (say 5) on the basis of the confounder score, and a stratified analysis or a logistic regression can be performed.

In the field of clinical trials Tukey (1991) describes how composites can be formed by weighing potential covariates on the basis of the degree of significance of individual univariate analyses between the covariate and the outcome (on treatment and control groups combined). So, for example, a variable for which the association with the outcome has a p-value$<$0.00002 could be scored 4, one with p$<$0.001 scored 3. The p-values are not used as indicators of significance but rather as indicators of 'more' or 'less'. He also

outlines a smear and sweep approach which can be used if the independent variables are categorical, and the data set has many observations. Two-way tables are formed of successive possible covariates. The outcome variable and the independent variable of primary interest are thus not used in the cross-classification of the table. For each cell of the table the outcome rate (say the death rate if the outcome is death or survival) in the exposed group (the independent variable of prime interest) is calculated and adjusted. To form new categories these adjusted death rates are ordered and "swept up" into the new categories, so that there are about equal numbers of the outcome in each new category. The apparent effect of the independent variable of prime interest is assessed by a two-way table consisting of the final categorisation and the independent variable of prime interest. Questions remain as to which variables should be entered into the smear and sweep procedure, in which order they should be considered and whether variables should be entered more than once.

4.3.7 Minimise the error of estimation (Schall and Zucchini 1990)

Schall and Zucchini (1990) propose a model selection approach which determines whether an extraneous factor should be treated as an effect modifier, confounder, or neither, so as to improve the accuracy of the estimate of the odds ratio of interest. The aim is thus not to determine whether the extraneous factor is in fact a confounder or an effect modifier. The model selected is the one which is estimated to maximise the accuracy of the estimator of the odds ratio of interest, on average. They make use of a non-parametric bootstrap method to carry out the selection. In their example an operating model is chosen in which the possible confounder is taken to be an effect modifier. All simplifications of this full model are taken to be approximating models. A discrepancy (measure of lack-of-fit) between the estimate of the operating model and that of approximating models is defined. The approximating model which, in a bootstrap sample, minimises the expected

discrepancy is chosen as the model on which to base effect estimation.

4.3.8 Partial Gauss discrepancy (Schall 1989)

Schall (1989) proposes the partial Gauss discrepancy for variable selection in linear models when the estimation of a subset of the parameters in the model is of interest.

The general linear model is denoted by

$$y = (X : \quad Z_1 \quad Z_2) \begin{pmatrix} \beta^* \\ \gamma_1^* \\ \gamma_2^* \end{pmatrix} + e$$

where estimation of β^* is of primary interest. To select that subset of variables of Z which leads to the most precise estimate of β^* the model with the smallest partial Gauss discrepancy is selected. The discrepancy

$$\Delta(\beta) = (\beta - \beta^*)' M (\beta - \beta^*)$$

is proposed where M is a nonnegative definite matrix. If $M = X'X - X'Z(Z'Z)^{-1}Z'X$, the criterion, an estimate of the expected overall discrepancy of the approximating model containing X and Z_1 becomes

$$\hat{\gamma}_2 Z_2' \tilde{X} (\tilde{X}'\tilde{X})^{-1} M (\tilde{X}'\tilde{X})^{-1} \tilde{X}' Z_2 \hat{\gamma}_2 - \hat{\sigma}^2 q$$

$$+2\hat{\sigma}^2 trace[(\tilde{X}'\tilde{X})^{-1} M]$$

where $\tilde{X} = (I - Z_1(Z_1'Z_1)^{-1}Z_1')X$. $\hat{\gamma}_2$ is obtained from the full model where Z_2 consists of the variables excluded from this approximating model. q is the number of parameters that are of prime interest.

This criterion can be adapted for the case of logistic regression by including the matrix W, which has diagonal elements $w_i = p_i(1 - p_i)$. The term $\hat{\sigma}$ equals one. The criterion then becomes

$$\hat{\gamma}_2 Z_2' W \tilde{X} W (\tilde{X}' W \tilde{X})^{-1} M (\tilde{X}' W \tilde{X})^{-1} \tilde{X}' W Z_2 \hat{\gamma}_2 - q$$

$$+2 trace[(\tilde{X}' W \tilde{X})^{-1} M]$$

where $\tilde{X} = (I - Z_1 W (Z_1' W Z_1)^{-1} Z_1' W) X$, and $M = X'WX - X'WZ(Z'WZ)^{-1}Z'WX$. For the full model, that is the model in which all confounders are fitted, Z_2 equals zero, and the criterion has the value of the number of parameters of interest.

4.4 Joint effects of factors lead to confounding

Hypothesis testing and the change-in-estimate criterion outlined above, are mostly applied in such a way that a variable is considered to be not a confounder if it is univariately found not to be a confounder. The joint effects of confounders are then investigated to see whether certain of the confounders can be excluded since they play no role when the other confounders are present (Miettinen 1974). Fisher and Patil (1974) propose, however, that one should not exclude a potential confounder from further consideration because it is found not to be a confounder on its own: two or more factors may jointly constitute a confounder. Miettinen (1974), however, states that the returns from such an approach would generally be low relative to the effort, and that adequate information of such conditional relationships would rarely be available.

4.5 Interactions between confounders

The interactions between confounders can also be considered as terms in the model, and effects of such interactions are generally assessed in the same way as confounders are assessed.

4.6 Categorising continuous confounders (Becher 1992)

Becher (1992) discusses the effect that the categorisation of a continuous confounder into a number of categories has on the successful removal of confounding. He shows in the case of logistic regression that categorisation into two levels may lead to the effect of the confounder not being removed completely, as Cochran (1968) has shown in the case of linear regression.

Categorising a continuous confounder into 5 or 6 levels is considered more appropriate.

If $\hat{OR}_c$ is the odds ratio of interest obtained when the confounder is included as a continuous variable, and $\hat{OR}_k$ the odds ratio of interest when the confounder is categorised into k levels, then the residual confounding effect of categorising the confounder into k categories is given by

$$\frac{\hat{OR}_k}{\hat{OR}_c}.$$

The relative residual confounding, which is a measure of the proportion of confounding which has been removed by the incomplete adjustment, assuming that full adjustment is achieved by including the variable as continuous, is given by

$$\frac{\ln\hat{OR}_n - \ln\hat{OR}_k}{\ln\hat{OR}_n - \ln\hat{OR}_c}$$

where $\hat{OR}_n$ is the odds ratio obtained when the confounder is not included in the model.

In this case the confounder as continuous variable is taken to reflect the true situation. Transformations of the continuous variable can also be used to reflect the true situation. "Positive" confounding is said to exist when the odds ratio is larger when the confounder is not adjusted for adequately; this leads to residual confounding larger than 1. "Negative" confounding, which exists if the odds ratio is smaller when the confounder is not adjusted for adequately, leads to residual confounding smaller than one. The interpretation of relative residual confounding is the same for both kinds of confounding.

Bootstrap samples can be taken to calculate confidence intervals for the residual confounding and relative residual confounding.

As mentioned in Section 3.6.2 it may be unrealistic to assume that the truth is reflected by the continuous variable, or even by a transformation of the continuous variable. Because of this it is common practice to categorise. The calculation of residual confounding and relative confounding may thus be meaningless, since the true relationship is unknown.

4.7 Selection of effect modifiers

4.7.1 Statistical criteria

The presence of effect modifiers (in statistical terms interaction between factors) is determined by the p-value for the test that the coefficient of the interaction term is zero (Hosmer and Lemeshow 1989) or by comparing the deviance of the model containing the interaction with that of the model without the interaction term (Collett 1991). Greenland (1983) shows that the latter approach has greater power.

Whereas statistical criteria are considered by many to be inappropriate for the selection of confounders, most authors propose that the decision whether effect modification is present (ie whether the interaction between the exposure of interest and an extraneous variable should be fitted in the model) should be based on statistical criteria (for example, Greenland 1989; Hosmer and Lemeshow 1989; Collett 1991). Walker (1986) points out that one generally has low power to detect interactions, but most authors feel that significance testing protects one against pursuing artefactual interactions (Fleiss 1986). Hosmer and Lemeshow (1989) add that the interaction should also be biologically plausible. Walker (1986) feels that external relevant observations may be crucial in deciding whether an observed interaction effect should be taken seriously. Walker (1986) and Thompson (1987) state that if subgroup analyses are of interest rather than interaction as such, there is no need to formally assess the interaction. However, if one wants to investigate interaction as such, Thompson (1987) and Greenland (1983) propose that this should be done by calculating a confidence interval for the interaction parameter.

4.7.2 Change-in-estimate criterion

The change-in-estimate criterion, as outlined above for detecting the presence of a confounder, is not appropriate for the detection of effect modification. As was outlined in Section 2.5 the inclusion of an interaction term

will produce changes in the estimated exposure coefficient, whether the interaction is significant or not, since the meaning of that coefficient is changed when interaction terms are included.

4.8 Summary of selection procedures

As in the case of approaches to selection of predictors, the approaches to selection of confounders and effect modifiers differ with respect to computational and intellectual complexity. On the intellectual side most epidemiologists stress that a priori knowledge must be used to determine which factors are relevant confounders to adjust for. It is, however, questionable whether such knowledge always exists. In addition, there may be a large number of possible confounders which in some way has to be brought down to a manageable number so as to obtain fairly precise estimates of effect. Table 4.2 gives a summary of the complexity of the procedures, whether there are criticisms against their use, and how often they are used in the medical literature.

Table 4.2: Summary of selection procedures for estimation

Procedure	Computationally	Intellectually	Criticisms	Used
significance testing	easy	easy	many	often
fit all confounders	easy	easy	many	often
prior beliefs	easy/difficult	difficult	few	seldom
change in estimate	easy	easy/difficult	few	often
composites	easy	easy	none	seldom
minimise estimation error	difficult	easy/difficult	none	never

The proposal by Schall and Zucchini (1990) is in a way completely different from all others in that they are not interested in the debate whether a factor is a confounder, an effect modifier or neither, but rather how the

factor should be considered in the analysis so as to obtain the most precise estimate of the effect measure of interest.

Significance testing is a clearcut, easily describable method, which is useful especially when little prior knowledge is available about confounders. On the negative side, sample size plays a big role in determining whether a factor is significantly associated with the outcome and/or exposure, and therefore it is suggested that a p-value of 0.25 is used, rather than the more conservative 0.05.

To fit all known confounders is computationally and intellectually undemanding, and is the method used most often in the medical literature (in 16 of the 30 articles and briefs published in the American Journal of Public Health in 1992 using logistic regression for estimation "all possible confounders were fitted"). However, sample size determines how many factors can reasonably be fitted, and estimates can be very imprecise if many confounders are fitted in a small data set, an issue not addressed in medical publications.

The change-in-estimate criterion is easily describable, but the random variability of different estimates is generally not taken into account.

Composites of possible confounders are seldom seen in the medical literature.

Minimising the error of estimation is computationally challenging, but selects the model which gives the most precise estimate of the effect measure of interest, the main goal of the analysis.

As far as selection of effect modifiers is concerned there seems to be agreement that significance testing is the only approach.

More specific comments on the various approaches will be given in Chapter 6 where the approaches will be tried out on various data sets.

CHAPTER 5: STANDARD STATISTICAL PACKAGES

In this chapter the logistic regression capabilities of the computer programmes SAS and BMDP will be described.

5.1 SAS

5.1.1 PROC LOGISTIC (SAS Institute Inc 1990)

This procedure fits linear logistic regression for a binary or ordinal outcome variable. The binary response data can be input in the form of count data from a binomial experiment. The independent variables must be numeric. By default, the programme uses the smallest value of the response variable as the outcome of interest (if the ORDER option is not used). If success (the outcome of interest) is coded as 1, failure should rather be coded as 2, not the usual 0.

Maximum likelihood estimates are computed using the iteratively reweighted least squares algorithm.

Model selection can be either forward, backward or stepwise. Variables can be forced to remain in the model. When backward or stepwise selection is used a value must be given to specify the significance level for staying in the model. By default this value is 0.05, ie the least significant variable with p-value>0.05 will be removed from the model. If forward or stepwise selection is used a value must be given to specify the significance level for entry into the model. By default this level is 0.05, ie the variable with the smallest p-value<0.05 is entered into the model. It can be specified how many times a variable can be entered or removed from the model.

For each model, the parameter estimates, their standard errors and odds ratios, Wald chi-square test statistic and p-value are printed. The model fit can be assessed by -2 log likelihood, the Akaike information criterion and the Schwarz criterion. A p-value is printed for the former statistic which has a chi-square distribution. The latter two criteria adjust the log likelihood criterion for the number of terms in the model and the number of observations

used. Akaike's information criterion is defined in the manual (SAS Institute Inc 1990) as

$$-2\ln L + 2(k+s)$$

and Schwarz's criterion as

$$-2\ln L + (k+s)\ln N$$

where k is the number of levels of the response variable, s the number of explanatory variables, and N the number of observations. However, these criteria seem to rather be defined as

$$-2\ln L + 2p$$

and

$$-2\ln L + p\ln N.$$

The manual suggests that these criteria should be used when comparing different models for the same data, with lower values indicating more desirable models.

The predictive ability of the model can be assessed by rank correlations between the observed outcome and the predicted probabilities, by classifying each pair of observations as concordant (discordant) if the outcome of success has a higher (lower) predicted probability of success than the failure outcome. Four indices of rank correlation are available: c, Somers' D, Goodman-Kruskal Gamma and Kendall's Tau-a.

A 2x2 frequency table of observed and predicted responses can be requested. A probability level has to be specified such that an observation with predicted probability greater than or equal to this level is classified as an event. The default value of this probability is 0.5. The sensitivity, specificity, false positive and false negative rates are calculated. A one-step approximation to the jackknife is used to reduce the bias caused by classifying the data from which the classification criterion was derived. In this

way each observation is deleted in turn so that the prediction of an observation's outcome is based on the prediction equation calculated using all other observations.

Various regression diagnostics are available to identify extreme points, observations not well explained by the model and observations which cause instability in the coefficients. Plots of these can also be requested.

An output data set containing, for example, predicted probabilities of success for each observation can be created. This can be used in a programme such as PROC REG to perform best subsets linear regression as outlined in Section 3.2.3.

The major drawback of this programme is that categorical variables cannot be incorporated into the analysis satisfactorily. For categorical variables with more than 2 categories one has to create design variables but there is no way in which one can inform the procedure that this set of design variables has to be considered as one set for entry or removal from the model. Similarly interaction terms have to be created as new variables and one cannot force the programme to include the interaction terms in the model only if the main effects have already been included.

As will be outlined below the procedure PROC CATMOD can be used to do logistic regression when the independent variables are categorical but this is not a very satisfactory approach either.

5.1.2 PROC CATMOD (SAS Institute Inc 1990)

This procedure was written to perform categorical data analysis, among others logistic regression. This procedure is more appropriate to use if some of the independent variables are categorical, but not advisable if some of the continuous independent variables have a large number of unique values. In that case PROC LOGISTIC is the programme of choice. Model selection is not performed in PROC CATMOD. The log-likelihood of a fitted model is printed and predicted probabilities can be requested.

5.1.3 PROC LOGIST (SAS Institute Inc 1986)

PROC LOGIST is only available in SAS Version 5 and seems to be the earlier version of PROC LOGISTIC which has only become available in SAS Version 6 (1990). The manual notes on PROC LOGIST contain various comments on model selection which are not included in the later version. Caution is given about stepwise variable selection and the number of variables to be entered in relation to the number of observations present. Furthermore, the manual advises against using stepwise variable selection to find significant confounding variables when one wants to test for a treatment effect after adjusting for other factors. All confounders should rather be included in the model along with the treatment variable. It is also pointed out that Akaike's information criteria should be used to judge which of a number of prespecified models has the best predictive ability. Models are not prespecified when stepwise variable selection is used. It is proposed that in variable selection one should use Akaike's information criterion and stop the selection process when the residual chi-square falls below twice its degrees of freedom.

5.2 BMDP (1983)

5.2.1 PLR: Stepwise logistic regression

PLR performs stepwise logistic regression where the outcome variable is coded as 0 or 1 (or is given as a count of successes or failures), and the independent variables are either continuous or categorical. In the model statement the user has to specify a continuous variable as such, since variables are otherwise assumed to be categorical. The programme creates design variables for all independent categorical variables and their interactions, for example, 2 design variables would be created for a categorical variable with 3 categories. By default "marginal" design variables are created such that each design variable used without the others contrast a category with the first, for example the two design variables reflecting three categories would

be -1 -1, 1 0, 0 1. However, "partial" design variables can be requested whereby each design variable used with the others contrasts a category with the first. For example, in the case of a variable with 3 categories the 2 partial design variables would take on the values 0 0, 1 0, 0 1. This enables one to easily calculate the odds ratio of a category compared to the first, using the estimated coefficients. Maximum likelihood estimates are obtained iteratively.

The whole set of design variables corresponding to a given categorical variable will be considered for inclusion/exclusion from the model at a given step. A hierarchical rule allows that a higher order interaction is only allowed in the model if all lower level interactions and main effects are included in the model. This rule can, however, be negated.

For each independent variable one can specify whether it must be in or out of the model at the start of the stepping procedure. In this way one can start with the full model or with the model containing no variables. One can also force certain important confounders to remain in the model since one can specify the number of times each variable is allowed to move in or out of the model.

At each step, selection is based on either the maximum likelihood ratio (MLR) or the approximate asymptotic covariance estimate (ACE). MLR is more reliable but ACE is faster and more economical and the manual (1983) recommends that that this is used in the initial run of a large problem. At each step the statistics for entry or removal of each variable are printed: F-statistics in the case of ACE, chi-square in the case of MLR. At each step the log-likelihood, the improvement chi-square based on the change in log-likelihood from the previous step, and three goodness-of-fit statistics are printed. A small p-value for the improvement chi-square indicates a significant improvement in that step.

The goodness-of-fit chi-square is based on the observed versus perdicted frequencies for each cell and thus tests whether the model at that step fits

the data adequately. In the manual (1983) it is pointed out that this test can be misleading if there are small cell frequencies. The Hosmer goodness-of -fit test is based on the observed and predicted frequencies after splitting the predicted values into 10 cells. A small p-value means that the model does not fit the data well. The CC Brown goodness-of-fit test compares the fit of the data to the logistic and a small p-value here would indicate that the logistic model is not appropriate for the data.

At each step the coefficients of the variables in the model are printed, with their standard errors, and odds ratio. The default p-value to leave the model is 0.15, and the default p-value to enter is 0.10.

At the end of the stepping procedure the frequencies of successes, failures, predicted probability, observed proportion, standardised residuals and log odds are printed for each distinct combination of the independent variables. Scatter plots of the observed proportions versus the predicted probabilities as well as of the observed proportions versus the the log odds, and histograms of the predicted probabilities for each group can be requested. A table of correct and incorrect classifications of the model can be requested, using various cutpoints to classify the predicted probabilities as sucesses/failures.

An output data set containing, for example, predicted probabilities can be created for use in other programmes.

CHAPTER 6: APPLICATIONS

In this chapter various selection procedures are applied to data sets. The procedures are compared with regard to the similarity of results obtained, the computational and intellectual complexity and the appropriateness of the procedure for the specific aim of the analysis. Practical decisions which have to be made when applying the procedures, and difficulties encountered, will be described. Based on the above, recommendations will be made about the choice of selection procedures.

6.1 Prediction

6.1.1 Example 1: Predictors of impaired glucose tolerance

Description of the problem

This example is used to illustrate the usefulness of various selection procedures when there is a fairly large number of possible predictors to select from in a data set with many observations. In addition, by taking a subsample of the data set, it is investigated whether sample size has an effect on the performance of the procedures.

In 1989 a survey covering various cardiovascular risk factors was conducted on a random sample of 854 adult (25 years and older) inhabitants of QwaQwa (Mollentze et al 1994). Since the researcher is particularly interested in diabetes, one of the analyses was aimed at identifying predictors of impaired glucose tolerance. Respondents who are on treatment for diabetes or have glucose level above 7.8 were categorised as having impaired glucose tolerance.

The researcher identified 11 possible predictors from amongst the wide-ranging questions on the questionnaire. All variables were categorised in 2 categories, namely a high risk (coded as 1) versus a low risk group (coded as 0), except for hypertension and smoking, which were categorised into 3 groups. It may seem somewhat crude to categorise all predictor variables

but as Van Houwelingen and Le Cessie (1990) point out, clinicians tend to categorise variables since there are clearcut risk groups, the associations between categorised variables and the outcome are easier to understand than those between continuous variables and the outcome, and the results are easier to apply in the clinical setting. The 11 possible predictors and their categories and coding are listed in Table 6.1. The researcher could not specify whether there might be any interactions that are biologically plausible. Of the study participants 793 had complete information for these variables as well as the outcome. Of these 793 participants 132 were classified as having impaired glucose tolerance.

Stepwise selection

Since the number of possible predictors is relatively large some stepwise selection would be the easiest analysis (computationally and intellectually) to attempt. The stepwise procedure of BMDP PLR was used with p-value to enter 0.10 and p-value to leave 0.15 (the defaults of BMDP), starting with the full model. Starting with the full model safeguards against the omission of a variable which is only significant after others have been fitted (Hosmer and Lemeshow 1989). The variables retained in the final model were age group, upper body obesity, cholesterol, ggt and sum of insulins. The coefficients, odds ratios and 95% confidence intervals for the odds ratios of this model are listed in Table 6.2. This model has log-likelihood of $-$ 313.908. Starting the selection procedure with the model containing only the intercept term led to the same final model.

Among the 11 predictors there are no variables which measure exactly the same characteristic as some other variable. Obesity and upper body obesity may seem to be measuring similar characteristics, but in the medical literature there is evidence that upper body obesity (ie apple shaped build compared to pear shaped build) is a risk factor for various diseases, whereas general obesity is not (Prof WF Mollentze, personal communication). Simi-

larly fasting insulin and the sum of insulins do not measure the same quantity. It would therefore not be expected that the approach of Hauck and Miike (1991) to determine close alternatives would be able to identify close alternatives. In the forward selection procedure, the p-value to enter of obesity went from 0.34 to 0.97 at the step where upper body obesity entered. The p-values associated with the variables at each step are listed in Table 6.3.

Bootstrap replications using stepwise selection

To investigate how often the model selected by stepwise selection procedures would be selected in resamplings, 300 random samples of 793 were drawn with replacement, using SAS/IML. For each sample, stepwise selection, starting with the model containing all 11 predictors, and with p-value to enter 0.10 and p-value to leave of 0.15 was performed, using SAS PROC LOGISTIC. Since this programme cannot be instructed to treat the two dummy variables of a 3-level categorical variable as a set for inclusion or exclusion from the model, the two 3-level variables, smoking and hypertension, were recategorised into two categories. Smoking was categorised as ever smoked (coded 1) and never smoked (coded 0), whereas hypertension was coded as hypertensive (high blood pressure or on hypertensive treatment, coded as 1) and normotensive (normal blood pressure and not on treatment, coded as 0). (Using backward elimination on the full data set including these 2-level variables rather than the 3-level ones used before, led to the same final model as in Table 6.2).

It was determined how often each variable was retained in the final model selected, and these results appear in Table 6.4. The variables which appeared in Table 6.2 are the ones which were selected most frequently, except for insulin sum, which appeared less frequently than fasting insulin. The model selected in Table 6.2 was selected as final model in only 6 of the 300 bootstrap samples. The model selected most frequently consisted of age group, upper body obesity, ggt and fasting insulin, and was selected as final model in 15

of the 300 bootstrap samples. The combination of age group, upper body obesity and cholesterol appeared in 151 of the final model selected. Although the bootstrap samples lead to varying models being selected, it does seem clear that certain variables appear frequently, whereas others (for example smoking and hypertension) appear less frequently. (This programme took a few hours to run on a 386 PC with 8M.)

Fitting all predictors

Table 6.5 lists the coefficients of the full model containing all 11 predictors. This model has log-likelihood of -309.154. As can be seen the odds ratios in the smaller model (Table 6.2) do not differ much from the odds ratios in the full model and the confidence intervals are not wider in the full model than in the smaller model. Of the predictors included in the smaller model, ggt and sum of insulins fail to reach significance at the 5% level in the full model, but have $p<0.10$. Fitting all possible predictors and deciding which variables are significant, based on the Wald statistics in the full model thus does not lead to conclusions which differ markedly from those reached when using a stepwise selection procedure.

Akaike's information criterion

To determine which models have small values for Akaike's information criterion, all possible models should be fitted. In the case of 11 possible predictors this would entail fitting 2047 models. Since this is an extremely large number of models to fit it was decided to eliminate from consideration those variables which on the the univariate analysis had a chi-square p-value of greater than 0.25. In this way sex, obesity and smoking were eliminated, leaving 8 predictors to consider, ie 225 models to be fitted. (If $p>0.10$ was used as cutpoint fasting insulin would have been excluded in addition. All other variables have $p<0.05$.) These models were fitted using SAS PROC LOGISTIC. Two dummy variables were created for the 3-level variable hypertension, using reference cell coding. The predictors included in the 10

models with the smallest value for Akaike's information criterion are listed in Table 6.6. The model with the smallest value is thus once again the model containing age group, upper body obesity, cholesterol, ggt and the sum of insulins. There are, however, many other models with fairly similar values for Akaike's criterion. It should be noted that age group, upper body obesity and cholesterol are common to all these models. The model containing only these three variables has AIC=644.63. Models excluding these three variables all have AIC of approximately 700 and higher. The model with the lowest value for Schwarz's criterion contains age group, cholesterol and upper body obesity.

Purposive selection

To follow the purposive selection procedure advocated by Hosmer and Lemeshow (1989), one would fit all variables as the first step. This was done with the 11 possible predictors. The variables with large p-values for the Wald statistic (hypertension, fasting insulin and sex) were eliminated, and the smaller model compared with the full model by the likelihood ratio test. The difference in deviance was 621.890 - 618.309 = 3.581 with 4 degrees of freedom, thus indicating that the eliminated variables did not improve the fit. Furthermore, the coefficients of the retained variables were similar in the smaller and the full model, indicating that the excluded variables did not need to be included to make adjustments to the coefficients. The researcher had specified that the variables considered were all biologically plausible predictors, and felt that the odds ratios of the variables retained in the model "made sense".

Best subsets linear regression

To use Hosmer and Lemeshow's approach of best subsets linear regression, the 8 predictors identified as having univariate association with outcome of $p<0.25$ were fitted as the full model in SAS PROC LOGISTIC, and the predicted probabilities were output to an output data set, in which the new

dependent variable and the case weights were defined as new variables. The SAS commands were as follows

```
PROC LOGISTIC;
MODEL DIABGRP=AGEGRP2 WHRATIOM HYP1 HYP2 TRIGHI
CHOLHI GGTHIGH INSFGRP INSSGRP;
OUTPUT OUT=OUT P=PROB;
DATA OUT;SET OUT ;
Z=
LOG(PROB/(1-PROB)) + (DIABMELL - PROB)/(PROB*(1-PROB));
W= PROB*(1-PROB);
```

(The variable DIABGRP is coded as: 1=present, 2=absent, whereas the variable DIABMELL is coded as 1=present, 0=absent. DIABGRP was used as dependent variable in PROC LOGISTIC, since the programme reads the smallest code of the outcome variable as "present".)

SAS PROC REG was used to do best subsets selection, using the new dependent variable and the defined case weights. Since 2 dummy variables had to be created to deal with the 3-level variable hypertension, and there is no way in which one can inform SAS that these variables must either both be included or both be excluded from a model, the best subsets selection identified models containing only one of the two dummy variables, and these models have to be ignored. The models with C_q approximately equal to the number of predictors fitted plus the intercept term, are listed in Table 6.7.

Small sample

The selection approaches used on the above data set all identified a similar set of predictors as being important. To see whether the different approaches would give such a stable response in a smaller sample (for example, does fitting all predictors give the same results as the selection procedures), a simple random sample of 150 was drawn out of the 793 cases. Of the 150

participants 28 had impaired glucose tolerance. Clearly, one would not expect that the same variables which were selected in the full data set would also be selected in the sample.

The results of the stepwise selection, starting with the full model, using BMDP PLR are shown in Table 6.8, and the results of fitting the full model in Table 6.9. The estimates of the odds ratios differ somewhat when comparing the full model with the smaller model. Ggt has p=0.05 in the smaller model compared to p=0.18 in the full model. Starting the stepwise selection with the model containing only the intercept term led to the identification of a different model from the final model obtained when starting the stepwise selection with the full model, and these results are given in Table 6.10. The p-values associated with the variables at each step are listed in Table 6.11. The log-likelihood of the model in Table 6.8 was -56.389, in Table 6.9 -53.694 and in Table 6.10 -60.177.

Before models were fitted to compare Akaike's information criterion, predictors with univariate chi-square p-value greater than 0.25 were eliminated. Three variables were eliminated in this way, leaving age group, sex, obesity, upper body obesity, hypertension, cholesterol, triglycerides and ggt as possible predictors. (If variables with p>0.10 were eliminated obesity, sex and cholesterol would have been excluded. All remaining variables have p<0.05.) Using SAS PROC LOGISTIC all 255 models were fitted. The 10 models with smallest values for AIC are listed in Table 6.12.

Table 6.13 gives the models identified by best subsets linear regression as having small C_q values (ignoring models which contained only one of the dummy variables associated with hypertension).

Using the purposive selection procedure of Hosmer and Lemeshow, the 8 selected predictors were fitted, and the variables with large p-values for the Wald test eliminated. Only age, hypertension and body mass index were retained, with a likelihood ratio test of 8 with 5 degrees of freedom.

Discussion

The small sample results indicate that the coefficients and p-values obtained when fitting all variables can differ markedly from those obtained when fitting a smaller selected model. The practice of fitting all possible predictors and deciding on their significance based on the full model should thus be discouraged. The small sample results also show that various selection procedures identify different models as being important. This indicates that there is no best model, and that the possible alternatives should be mentioned.

In the full data set the various selection procedures identified similar models as being important. However, only by trying out various selection procedures and thus determining that similar sets of predictors are selected by all procedures, can one confidently say that an important core has been identified. Furthermore, although there are no predictors which are clear proxies for one another in this data set, there are various models which fare similarly well with regards to AIC. It must thus be stressed to the researcher that different models can perform equally well, even if the variables in the one are not simply substitutes for variables in the other.

Fitting all models to compare them on the basis of some criterion, for example Akaike's information criterion, becomes cumbersome even with only 11 possible predictors. Some pre-selection based on a univariate assessment makes this approach more practical.

Issues regarding categorisation of continuous predictors, and interactions between predictors will be dealt with in the next example.

6.1.2 Example 2: Demographic predictors of smoking in females

Description of the problem

This example is used to illustrate the usefulness of various selection procedures in a large data set with few possible predictors to select from. In a survey of all inhabitants of a town with 4000 inhabitants near Cape Town,

a wide range of health information was collected (Hoffman et al 1988). A subanalysis dealt with determining whether age, education and employment status were predictors of smoking among women aged 25 to 64 years (Yach and Joubert 1988). The categorisation and coding of variables are listed in Table 6.14. The researcher indicated that there may be interactions between age and education, and age and employment. The information of 531 women could be used in the analysis, with 266 of them being smokers.

Stepwise selection

As a first analysis, stepwise selection starting with the full model was done using BMDP PLR, fitting all main effects, as well as the interactions between age and education, and age and employment. The final model selected consisted only of the main effects as listed in Table 6.15. This model has log-likelihood of -351.610. Starting the selection with only the intercept term in the model led to the same final model.

Fitting the full model

Fitting the full model containing the possible interactions, led to the coefficients listed in Table 6.16. This model has log-likelihood of -350.807.

Akaike's and Schwarz's information criteria

Table 6.17 lists the AIC and SIC of all possible models. The model with smallest AIC is the one containing all three predictors, but no interactions. The model with smallest value for SIC is the model containing age and education.

Error rates

The apparent counting error rates for each model, using a cutoff of 0.5 for the predicted probability of smoking to define a smoker, are listed in Table 6.18. The error rates are generally high, which might be due to the fact that the cutoff of 0.50 is inappropriate. The models including interaction terms

have slightly lower error rates. The cross-validation error rates were found to be identical to the apparent error rates, since eliminating one observation at a time from such a large data set can have little influence on the model. Surprisingly the bootstrap error rates also provided little correction to the apparent error rates. In some cases the bootstrap error rates are in fact smaller than the apparent error rates, which may be due to the large sample size, and the apparent error rates thus not being biased.

Best subsets linear regression

To use all subsets regression in SAS all interaction terms had to be created as new variables. Using the full model as starting point for the all subsets regression, leads to many models which contain only interaction terms and no main effects to be selected as models with small C_q's. Many models selected by the best subsets procedure thus have to be ignored. Table 6.19 lists the viable models which had C_q values close to the number of parameters fitted.

Age as continuous predictor

In the above analyses the predictor age was dichotomised as requested by the researcher. To determine whether any different conclusions would have been drawn had age not been categorised, some analyses were repeated.

Age as continuous variable remains a significant predictor and the models identified above as having lowest values for AIC and SIC remain the models with lowest values.

To investigate the need to transform the continuous predictor, age quartiles were created. Table 6.20 outlines the midpoint of each quartile. Three dummy variables were created, taking the lowest quartile as the reference group. The coefficients, odds ratios and 95% confidence intervals for the odds ratios are also indicated in Table 6.20. The coefficients of the first three age quartiles seem to follow a linear trend, with the fourth quartile falling out of line, possibly indicating the need for a quadratic term. Alternatively one

could also consider grouping the variable differently since the last two quartiles have similar odds ratios. Similar results were obtained when education and employment were also included in the model.

The model containing only age has deviance of 711.724 which decreases to 711.527 when the variable agelog(age) is added to the model. Similarly the inclusion of the variable age^2 leads to a decrease in deviance of 0.207. This would seem to indicate that there is no need to transform the variable age.

Discussion

In this example, since there are only a few predictors to consider, the selection procedures which require one to fit all possible models, are not as arduous as in the previous example. The model chosen by most procedures is the one containing all the main effects. Not only are the interaction terms not significant, but Akaike's information criterion is smallest for the model with only the main effects. As far as the error rates are concerned, the inclusion of the interaction terms leads to a very slight improvement. The cross-validation and bootstrap error rates were virtually identical to the counting error rate. In such a large data set it is to be expected that the exclusion of one observation for its own prediction (ie for cross-validation) would have no effect. The large sample size may also imply that the apparent error rates are not biased. Using the best subsets approach is a nightmare since one has to sift through the listed models to find ones which do not only consist of incomplete sets of dummy variables or interaction terms.

6.2 Estimation

6.2.1 Example 1: Estimating the effect of obesity on hypertension

Description of the problem

This example is used to illustrate the usefulness of various selection procedures when the aim of the analysis is estimation. A large data set with

a few possible confounders is used.

In the data set described in Section 6.1.1 the researcher wished to estimate the strength of the association between obesity and hypertension. Obesity was defined as in Section 6.1.1 but hypertension was categorised in 2 groups: hypertensive (diastolic 95 or above, or systolic 160 or above, or on hypertension treatment) versus normotensive. Age and sex could be possible confounders, and there may be interactions involved. Age was categorised in 5 age groups (25 to 34, 35 to 44, 45 to 54, 55 to 64, 65 and older) thus needing 4 dummy variables. Model selection thus had the aim of selecting the model giving the best estimate of the odds ratio for hypertension, associated with obesity. Information was available for 850 individuals of whom 303 were hypertensives. The researcher could identify which variables are possible confounders, but could not specify the expected size of the effect of the confounders.

Confounders associated with outcome and exposure

Age and sex are classical confounders in the data set, since both variables are highly significantly associated with obesity as well as hypertension. The odds of a female being hypertensive is 1.64 times that of a male (95% CI 1.20 to 2.23). The odds of a female being obese is 4.33 times that of a male (95% CI 3.18 to 5.90). Table 6.21 indicates the prevalence of obesity and hypertension in each age group. Whereas hypertension increases with age, obesity is highest in the middle age groups and lowest in the oldest age group. When investigating the association between obesity and hypertension one would therefore expect that the omission of sex may introduce a spurious obesity effect, whereas the omission of age may mask an effect of obesity on hypertension.

Change-in-estimate criterion

To use the change-in-estimate criterion, obesity is fitted on its own, and then in combination with each of the confounders and their interaction, to

determine whether the estimate of the odds ratio of interest is effected by the inclusion of the possible confounders. As proposed by Mickey and Greenland (1989), a change of 10% in the estimate of the odds ratio is taken as evidence that the factor is a confounder. The various models fitted in this way and the odds ratios associated with obesity are indicated in Table 6.22. From this it seems that the inclusion of age with obesity changes the estimate, whereafter the inclusion of sex changes the estimate once more, but the inclusion of the interaction term thereafter does not alter the estimate. The variables to include are thus obesity, sex and age group.

Minimising the error of estimation

Following the approach of Schall and Zucchini (1989), the model containing obesity, sex, age and the interaction between sex and age was fitted as operating model. From this model the odds ratio of interest ($\tilde{\rho}$) was calculated and the predicted probability of the outcome hypertension determined for each observation (p_j). A parametric bootstrap was performed by generating new observations from the binomial distribution $B(1, p_j)$. Each subset model containing obesity was fitted and the log odds ratio of interest computed. 300 samples were selected in this way, and for each model the mean squared error (MSE) was estimated as

$$\frac{1}{300}(\sum_{i=1}^{300}[\ln(\hat{\rho}_i) - \ln(\tilde{\rho})]^2)$$

The model with the smallest value for MSE was the model containing only obesity. The mean squared errors of all fitted models are indicated in Table 6.23.

A nonparametric bootstrap was also performed in which 300 random samples of 850 observations were drawn with replacement from the data set. The MSE of each model was estimated as before. These MSE's are also indicated in Table 6.23, and are in the same ranking order as those of the parametric bootstrap.

Partial Gauss discrepancy

Table 6.24 indicates the estimates of the partial Gauss discrepancy for the various models containing obesity. In this example the full model has criterion equal to 1, since the interest is focussed on one parameter. The model containing only obesity has the largest value for the criterion. These estimates were calculated using the predicted probabilities from the full model. If, however, the predicted probabilities from the model containing only obesity is used to calculate the weights, the estimates of the criterion change dramatically, as indicated in the second column of Table 6.24.

Categorisation of a continuous confounder

To determine the effect that categorisation of a continuous variable has on the removal of confounding, the variable of interest, obesity, and the confounder sex were fitted alone, as well as with age in various categorisations. In Table 6.25 the coefficients and odds ratios associated with obesity are indicated for the various models. Since transformations of age had little effect on the estimates of the odds ratio, the continuous variable (untransformed) is taken to present the truth. Relative residual confounding is thus calculated with respect to the model containing age as continuous variable. Categorisation into 5 categories gives very similar results to those obtained when age is continuous. However, dichotomising into various dichotomies leads to negative as well as positive confounding.

Contrast to prediction selection

If prediction selection procedures were used, different models would have been chosen. Both AIC and backwards elimination select as best predicting model the model containing obesity, sex, age and the interaction of sex and age. It is not at all viable using the best subsets approach since most of the models selected as having small C_q values include some but not all of the dummy variables.

Interactions

To evaluate possible interactions between obesity and the confounders (ie that the confounders are effect modifiers) the deviances of the various models were investigated. These are listed in Table 6.26. Removing the three-way interaction leads to a change of 3.4 with 3 degrees of freedom, removing the interaction between sex and obesity thereafter leads to a change of 0.7 with 1 degree of freedom, whereas removing the interaction between age and obesity leads to a change of 5.4 with 4 degrees of freedom. This would seem to indicate that there is no evidence that the two confounders are effect modifiers.

Discussion

This example shows how important it is to determine whether the aim of the analysis is prediction or estimation since prediction selection procedures lead to a different model than the model selected by confounder selection procedures. Some of the literature is unclear about this issue, with Day, Byar and Green (1980), for example, stating that a confounder is a variable which, in the presence of the exposure of interest, improves the prediction of the outcome.

Furthermore, including more terms in the model (in this case interactions between possible confounders) does not necessarily mean that a better estimate of the effect of interest has been found. The practice of including all possible confounders in the model should thus be discouraged.

The various confounder selection procedures did not all select the same model. The results using the partial Gauss discrepancy are confusing, since the choice of weights plays such a large role. In the case of predictor selection, the best subsets approach using the C_q criterion (Hosmer, Jovanovic and Lemeshow 1989) also makes use of weights based on the predicted probabilities of the full model. It was investigated on the data sets used in Sections 6.1.1 and 6.1.2 whether taking predicted probabilities from a smaller model

than the full one to obtain the weights would influence the values of C_q obtained. Different choices of weights made little difference. It is unclear why the effect of the weights should have such a large effect on the Gauss discrepancy. The usefulness of the partial Gauss discrepancy is thus not clear.

As in the case of prediction problems where it may be necessary to report various "best" models, it may be useful to report the various estimates of the effect of interest, obtained by various adjustments.

Table 6.1: Predictors of impaired glucose tolerance
Description of predictors

Variable name	Category	Code	
age group	25-44	0	
	45 and older	1	
sex	Male	0	
	Female	1	
obesity	Body mass index under 25	0	
	Body mass index 25 and above	1	
Upper body obesity	Below sex-specific median for waist to hip ratio	0	
	On or above sex-specific median for waist to hip ratio	1	
Fasting insulin	Fasting insulin on or below the 75th % of healthy reference group	0	
	Fasting insulin above the 75th % of healthy reference group	1	
Sum of insulin	sum of 1 and 2 hour insulin on or below 75th % of healthy reference group	0	
	sum of 1 and 2 hour insulin above 75th % of healthy reference group	1	
Cholesterol	below 5.2	0	
	5.2 and above	1	
Triglycerides	less than 2.3	0	
	2.3 and above	1	
Smoking	Never	0	0
	previously, not current	1	0
	current	0	1
GGT level	65 and below	0	
	above 65	1	
Hypertension	Diastolic below 95 and systolic below 160 and not on hypertension treatment	0	0
	Diastolic 95 or higher or systolic 160 or higher, but not on treatment	1	0
	On hypertension treatment	0	1

Table 6.2: Predictors of impaired glucose tolerance
Final model selected using backwards elimination

Variable	coefficient	Wald statistic p-value	Odds ratio	95% CI
age group (above 45 vs under 45)	1.1737	<0.01	3.23	(2.02; 5.17)
upper body obesity (above median to below)	0.9373	<0.01	2.55	(1.65; 3.95)
cholesterol (high versus normal)	0.63145	<0.01	1.88	(1.26; 2.80)
ggt (high versus normal)	0.57305	0.03	1.77	(1.06; 2.98)
insulin sum (high versus normal)	0.44173	0.03	1.56	(1.04; 2.32)

Table 6.3: Predictors of impaired glucose tolerance
P-values to enter at each step of the forward selection procedure

	Step 1	Step 2	Step 3	Step 4	Step 5	Step 6
age group	0.00*	0.00	0.00	0.00	0.00	0.00
upper body obesity	0.00	0.00*	0.00	0.00	0.00	0.00
obesity	0.35	0.34	0.97	0.50	0.73	0.40
sex	0.27	0.32	0.43	0.63	0.27	0.38
hypertension	0.00	0.05	0.09	0.09	0.12	0.17
smoking	0.27	0.40	0.37	0.44	0.24	0.29
triglycerides	0.00	0.00	0.02	0.05	0.11	0.17
cholesterol	0.00	0.00	0.00*	0.00	0.00	0.00
ggt	0.01	0.01	0.04	0.03*	0.03	0.03
fasting insulin	0.10	0.03	0.21	0.23	0.18	0.63
insulin sum	0.00	0.00	0.02	0.03	0.03*	0.03

*: Variable was entered at this step. P-values thereafter refer to p-values to leave the model.

Table 6.4: Predictors of impaired glucose tolerance
Variables appearing in final backwards elimination model in 300 bootstrap samples

	number of samples	%
age group	299	99.7%
upper body obesity	297	99.0%
obesity	105	35.0%
sex	101	33.7%
hypertension	84	28.0%
smoking	67	22.3%
triglycerides	102	34.0%
cholesterol	153	51.0%
ggt	287	95.7%
fasting insulin	204	68.0%
insulin sum	122	40.7%

Table 6.5: Predictors of impaired glucose tolerance
Fitting all possible predictors

Variable		coefficient	Wald statistic p-value	Odds ratio	95% CI
age group		1.1593	<0.01	3.12	(1.96; 5.19)
upper body obesity		0.94792	<0.01	2.58	(1.64; 4.06)
obesity		-0.35389	0.16	0.70	(0.43; 1.14)
sex		0.17807	0.53	1.19	(0.68; 2.10)
hypertension	(1)	-0.21213	0.39	0.81	(0.50; 1.32)
	(2)	0.35030	0.27	1.42	(0.77; 2.63)
smoking	(1)	-0.73692	0.16	0.48	(0.17; 1.24)
	(2)	-0.13102	0.67	0.88	(0.48; 1.61)
triglycerides		0.43873	0.17	1.55	(0.83; 2.91)
cholesterol		0.64215	<0.01	1.90	(1.25; 2.88)
ggt		0.55005	0.06	1.73	(0.98; 3.07)
fasting insulin		0.14441	0.56	1.16	(0.71; 1.89)
insulin sum		0.37247	0.10	1.45	(0.93; 2.27)

Table 6.6: Predictors of impaired glucose tolerance
Akaike's information criterion: the 10 models with smallest values

Variables included in the model	AIC
age group, upper body obesity, cholesterol, ggt, sum of insulins	639.815
age group, upper body obesity, triglycerides, cholesterol	641.392
age group, upper body obesity, triglycerides, cholesterol, ggt	642.076
age group, upper body obesity, hypertension, triglycerides, cholesterol, sum of insulins	642.106
age group, upper body obesity, cholesterol, sum of insulins	642.269
age group, upper body obesity, hypertension, cholesterol, insulin sum	642.441
age group, upper body obesity, cholesterol, ggt	642.454
age group, upper body obesity, hypertension, cholesterol, ggt	642.510
age group, upper body obesity, hypertension, triglycerides, cholesterol, ggt	642.625
age group, upper body obesity, cholesterol, ggt, fasting insulin	642.789

Table 6.7: Predictors of impaired glucose tolerance
Best subsets linear regression: models with Cp close to number of parameters

Variables included in model	number of variables	C_q
age group, upper body obesity, cholesterol, ggt and insulin sum	5	6.95
age group, upper body obesity, triglycerides, cholesterol, ggt, insulin sum	6	7.11
age group, upper body obesity, hypertension, cholesterol, ggt, insulin sum	7	7.50
age group, upper body obesity, hypertension, triglycerides, cholesterol, ggt, insulin sum	8	8.02
age group, upper body obesity, triglycerides, cholesterol, ggt, fasting insulin, insulin sum	7	9.03
age group, upper body obesity, hypertension, triglycerides, cholesterol, insulin sum	7	9.24
age group, upper body obesity, hypertension, cholesterol, ggt, fasting insulin, insulin sum	8	9.41
age group, upper body obesity, hypertension, triglycerides, cholesterol, ggt	7	9.68
age group, upper body obesity, hypertension, triglycerides, cholesterol, ggt, fasting insulin, insulin sum	9	10.0

Table 6.8: Predictors of impaired glucose tolerance: subsample
Final model selected using backwards elimination

Variable		coefficient	Wald statistic p-value	Odds ratio	95% CI
age group		2.3829	<0.01	10.84	(2.85; 41.27)
obesity		-0.92362	0.07	0.40	(0.14; 1.09)
hypertension	(1)	0.3644	0.50	1.44	(0.50; 4.18)
	(2)	1.9144	0.01	6.78	(1.71; 26.86)
ggt		1.2042	0.05	3.33	(1.02; 10.91)

Table 6.9: Predictors of impaired glucose tolerance: subsample
Fitting all possible predictors

Variable		coefficient	Wald statistic p-value	Odds ratio	95% CI
age group		2.3026	<0.01	10.00	(2.52; 39.76)
upper body obesity		0.62173	0.26	1.86	(0.62; 5.50)
obesity		-1.35459	0.04	0.26	(0.07; 0.92)
sex		-0.50089	0.47	0.61	(0.16; 2.37)
hypertension	(1)	0.36988	0.53	1.45	(0.45; 4.65)
	(2)	1.6332	0.04	5.12	(1.12; 23.38)
smoking	(1)	-0.78224	0.48	0.46	(0.05; 4.01)
	(2)	-0.02736	0.97	0.97	((0.21; 4.52)
triglycerides		0.31957	0.71	1.38	(0.26; 7.21)
cholesterol		0.68808	0.20	1.99	(0.69; 5.75)
ggt		0.92400	0.18	2.52	(0.65; 9.83)
fasting insulin		0.51367	0.46	1.67	(0.43; 6.55)
insulin sum		0.10733	0.87	1.11	(0.30; 4.12)

Table 6.10: Predictors of impaired glucose tolerance: subsample
Final model selected using forward selection

Variable	coefficient	Wald statistic p-value	Odds ratio	95% CI
age group	2.2549	<0.01	9.53	(2.70; 33.7)
triglycerides	1.4135	0.03	4.11	(1.12; 15.12)

Table 6.11: Predictors of impaired glucose tolerance: subsample
P-values to enter at each step of the forward selection procedure

	Step 1	Step 2	Step 3
age group	0.00*	0.00	0.00
upper body obesity	0.02	0.10	0.19
obesity	0.22	0.30	0.21
sex	0.20	0.24	0.35
hypertension	0.05	0.06	0.21
smoking	0.82	0.68	0.54
triglycerides	0.01	0.03*	0.03
cholesterol	0.18	0.26	0.45
ggt	0.04	0.06	0.16
fasting insulin	0.41	0.35	0.69
insulin sum	0.79	0.38	0.72

*: Variable was entered at this step. P-values thereafter refer to p-values to leave the model.

Table 6.12: Predictors of impaired glucose tolerance: subsample Akaike's information criterion

Variables included in the model	AIC
age group, hypertension, ggt, obesity	124.777
age group, upper body obesity, hypertension, ggt, obesity	125.241
age group, hypertension, cholesterol, ggt, body mass index	125.248
age group, upper body obesity, hypertension, obesity	125.414
age group, upper body obesity, hypertension, cholesterol, ggt, obesity	125.471
age group, upper body obesity, hypertension, cholesterol, obesity	125.586
age group, hypertension, triglycerides, ggt, obesity	125.884
age group, upper body obesity, hypertension, triglycerides, obesity	125.982
age group, upper body obesity, triglycerides, obesity	126.036
age group, hypertension, ggt	126.164

Table 6.13: Predictors of impaired glucose tolerance: subsample
Best subsets linear regression

Variables included in model	number of variables	C_q
age group, hypertension, ggt, obesity	5	6.36
age group, hypertension, cholesterol, ggt, obesity	6	6.83
age group, upper body obesity, hypertension, ggt, obesity	6	6.88
age group, upper body obesity, hypertension, cholesterol, ggt, obesity	7	6.99
age group, upper body obesity, hypertension, obesity	5	7.31
age group, upper body obesity, hypertension, cholesterol, obesity	6	7.37
age group, hypertension, triglycerides, ggt, obesity	6	7.47
age group, upper body obesity, hypertension, triglycerides, obesity	6	7.65
age group, hypertension, ggt, sex, obesity	6	7.83
age group, hypertension, cholesterol, ggt, sex, obesity	7	7.85
age group, upper body obesity, hypertension, triglycerides, ggt, obesity	7	8.07
age group, upper body obesity, hypertension, cholesterol, ggt, sex, obesity	8	8.22
age group, upper body obesity, hypertension, triglycerides, cholesterol, obesity	7	8.24
age group, upper body obesity, hypertension, cholesterol, sex, obesity	7	8.26
age group, hypertension, triglycerides, cholesterol, ggt, obesity	7	8.29
age group, upper body obesity, hypertension, ggt, sex, obesity	7	8.53
age group, upper body obesity, hypertension, triglycerides, cholesterol, ggt, obesity	8	8.55
age group, upper body obesity, hypertension, sex, obesity	6	8.71
age group, hypertension, cholesterol, ggt, sex	6	8.85
age group, hypertension, triglycerides, ggt, sex, obesity	7	9.16
age group, upper body obesity, hypertension, triglycerides, sex, obesity	7	9.39
age group, hypertension, triglycerides, cholesterol, ggt, sex, obesity	8	9.58

Variables included in model	number of variables	C_q
age group, upper body obesity, hypertension, triglycerides, cholesterol, sex, obesity	8	9.59
age group, upper body obesity, hypertension, triglycerides, ggt, sex, obesity	8	9.89
age group, upper body obesity, hypertension, triglycerides, cholesterol, ggt, sex, obesity	9	10.00

Table 6.14: Predictors of smoking
Description of predictors

Variable name	Category	Code	
age group	25-44	0	
	45-64	1	
education	Std 9 or higher	0	
	Below std 9	1	
employment	employed, non-manual	0	0
	employed, manual	1	0
	unemployed	0	1

Table 6.15: Predictors of smoking
Final model of backward elimination

Variable	coefficient	Wald statistic p-value	Odds ratio	95% CI
age group	0.88918	<0.01	2.43	1.45 to 4.09
employment (1)	0.050732	0.83	1.05	0.66 to 1.68
employment (2)	0.58491	0.03	1.80	1.06 to 3.02
education	1.0726	<0.01	2.92	1.60 to 5.34

Table 6.16: Predictors of smoking
Fitting all terms

Variable	coefficient	Wald statistic p-value
age group	1.1992	0.28
employment (1)	0.13976	0.78
employment (2)	1.9315	0.11
education	1.2465	0.27
age by employment (1)	-0.0926	0.87
age by employment (2)	-1.4284	0.25
age by education	-0.1951	0.87

Table 6.17: Predictors of smoking
Akaike's and Schwarz's information criteria

Variables entered in model	AIC	SIC
age	727.786	736.339
employment	734.732	747.562
education	726.725	735.278
age, employment	724.534	741.640
age, employment, age*employment	726.759	752.419
age, education	714.250	727.080
age, education, age*education	716.164	733.271
employment, education	723.059	740.165
age, employment, education	713.221	734.604
age, education, employment, age*employment	715.643	745.579
age, education, employment, age*education	715.134	740.794
age, education, employment, age*education, age*employment	717.614	751.827

Table 6.18: Predictors of smoking
Error rates

Variables entered in model	apparent	bootstrap
age	44.2%	44.1%
employment	45.5%	44.8%
education	44.7%	44.7%
age, employment	44.5%	43.0%
age, employment, age*employment	43.8%	42.6%
age, education	40.2%	40.5%
age, education, age*education	40.2%	40.5%
employment, education	43.8%	42.9%
age, employment, education	40.2%	40.1%
age, education, employment, age*employment	39.8%	40.0%
age, education, employment, age*education	40.2%	39.7%
age, education, employment, age*education, age*employment	39.8%	39.7%

Table 6.19: Predictors of smoking
Best subsets linear regression

Variables included in model	number of variables	C_q
age group, education, employment	4	3.73
age group, education, employment, age/education interaction	5	5.57
age, education, employment, age/employment interaction	6	6.05
age, education, age/education interaction	3	6.42
age, education, employment, age/education interaction age/employment interaction	7	8.00

Table 6.20: Predictors of smoking
Quartile analysis of age to investigate the need for transformation

Quartile	Midpoint	Number	Coefficient	Odds ratio	95% CI for odds ratio
1	27.3	132	0	1.00	
2	31.9	137	-0.289	0.75	0.46 to 1.22
3	37.5	134	-0.839	0.43	0.26 to 0.71
4	52.9	130	-1.116	0.33	0.20 to 0.54

Table 6.21: Estimating the effect of obesity on hypertension
Prevalence of obesity and hypertension by age group

Age group	group size	obesity		hypertension	
		n	%	n	%
25-34	197	99	50.25%	22	11.2%
35-44	178	105	59.0%	46	25.8%
45-54	126	75	59.5%	48	38.1%
55-64	147	93	63.3%	72	49.0%
65+	202	91	45.05%	115	56.9%

Table 6.22: Estimating the effect of obesity on hypertension
Odds ratios associated with obesity, in models excluding and including possible confounders

Variables included in model	OR associated with obesity	95% CI
obesity	2.34	1.75 to 3.14
obesity and sex	2.19	1.61 to 2.98
obesity and age	2.84	2.05 to 3.94
obesity, sex and age	2.54	1.80 to 3.58
obesity, sex and age and sex*age	2.57	1.82 to 3.64

Table 6.23: Estimating the effect of obesity on hypertension
Mean square error of 300 bootstrap samples

	Parametric bootstrap	Nonparametric bootstrap
Variables included in model	MSE	MSE
obesity	0.026	0.025
obesity and sex	0.045	0.040
obesity and age	0.040	0.039
obesity, sex and age	0.031	0.027
obesity, sex and age and sex*age	0.033	0.027

Table 6.24: Estimating the effect of obesity on hypertension
Partial Gauss discrepancy

Variables included in model	Using weights from full model Criterion	Using weights from model containing only obesity Criterion
obesity	2.16	0.88
obesity and sex	4.05	0.98
obesity and age	0.98	1.04
obesity, sex and age	0.99	0.99
obesity, sex and age and sex*age	1.00	1.00

Table 6.25: Estimating the effect of obesity on hypertension
Odds ratios associated with obesity, in models with the confounder age in various categorisations

method of adjustment for age	coefficient	OR	residual confounding	relative residual confounding
not adjusted for	0.784	2.19		
continuous	0.928	2.53	1.00	
5 categories	0.933	2.54	1.00	1.28
2 categories[a]	0.883	2.42	0.96	0.68
2 categories[b]	0.974	2.65	1.05	1.32
3 categories	0.975	2.65	1.05	1.33
square root of age	0.921	2.51		
log of age	0.908	2.48		

a: age dichotomised as less than 45, 45 and older
b: age dichotomised as less than 65, 65 and older

Table 6.26: Estimating the effect of obesity on hypertension
Deviances

Variables included	Deviance	parameters
A: three-way interaction	924.928	19
B: all two-way interactions	928.285	15
C: B- sex*obesity	928.954	14
D: C-age*obesity	933.837	10
E: B-age*obesity	933.671	11
F: E-sex*obesity	933.837	10

CHAPTER 7: RECOMMENDATIONS

Based on the literature review in Chapters 3 and 4 and the results obtained in Chapter 6, recommendations about the choice of selection procedures will be made in this chapter.

7.1 The role of the researcher

The analysis should be done in close consultation with the researcher, so as to obtain models which are biologically plausible, as well as useful to the researcher. In the case of prediction problems the researcher has a role to play regarding the following points:

- identification of variables which are close alternatives to one another. Suggestions should be made as to which of the alternatives should be considered for inclusion in the model
- identification of biologically sound interactions which should be evaluated
- categorisation of continuous predictors: models which contain continuous predictors or transformations of continuous predictors may be difficult to use practically. It may be more useful to the researcher, and more appropriate for the model, to investigate whether individuals in a certain risk group (for example, hypertensives) are more likely to have the outcome compared to those in the normal risk group, than to know that a certain risk is associated with a specific unit increase in a continuous variable (for example blood pressure).

In the case of estimation problems the researcher has a role to play regarding the following:

- identification of possible confounders and effect modifiers
- decisions as to what constitutes a clinically significant association and a clinically not significant association between the confounder and exposure/outcome
- providing prior information on the size of the effect of a confounder

- categorisation of continuous confounders

From my experience many researchers are not able to give valuable input into the model building process. It is our task to educate them that they indeed have an important role to play, and that the selection procedures cannot on their own do the best job. It sometimes seems as if statisticians have done such a good job of "selling" procedures such as stepwise selection, that researchers have far more confidence in them than we as statisticians do. As Achen (1982) expresses the problem: "statistical methods are simply tools, and one cannot use a tool well with no clear purpose in mind". "If prior theory and investigation suggested nothing ... one will tend to capitalize on chance, selecting variables and functional forms that do well in this one example".

7.2 The aim of model fitting

It is important to determine whether the aim of the analysis is estimation or prediction since different selection procedures are appropriate for each aim, and using an inappropriate approach may lead to wrong conclusions. If a researcher is primarily interested in the effect of a certain exposure, it is not useful to inform the researcher that the effect is not significant (for example in a stepwise selection procedure); rather one should estimate the effect and on the basis of the odds ratio and its confidence interval the researcher can decide whether the effect is clinically of no importance.

7.3 Prediction

As has been shown in the examples, fitting all predictors and determining their significance from the full model can (because of sample size) lead to different conclusions from those obtained from some selection procedure. If there are many possible predictors some selection should be done.

The various procedures available to the analyst differ with regard to their complexity. Guidelines which are practical are therefore needed to be able to make a choice between procedures.

Many criticisms have been levelled against stepwise selection procedures. Even the approach of Hauck and Miike (1991) is of limited use, since this approach identifies close alternatives, but not subsets of variables which could fare as well as the variables selected. Using bootstrap replications to see which variables are consistently selected in stepwise selection procedures, gives some safeguarding against inappropriately selecting one model as the best, but the problems of stepwise selection procedures still underlie this approach. An approach which forces one to assess all models should rather be used. To make such procedures practically feasible in data sets with many possible predictors a univariate screening (with say $p<0.25$) can be carried out initially. The most practical procedure is to calculate AIC or SIC for all models, and to report the models with the small values for these criteria. If there are many k-category variables for which dummy variables have to be created, or interaction terms, using best subsets linear regression becomes cumbersome, since one has to sift through many inappropriate models, namely models consisting of some but not all of the dummy variables relating to one variable, or interaction terms but not the main effects. The use of error rates as selection tools seems problematic since issues regarding the cutoff chosen to define a predicted outcome as an error, and the relative importance of various errors are problematic.

If possible, more than one approach should be used to determine whether there is a core of variables which can be identified as important. If different models perform similarly well, they should all be reported.

7.4 Estimation

Many of the procedures proposed in the literature rely heavily on prior knowledge and strong beliefs of clinicians. Since these often are not available one of the approaches tried on the estimation example in Chapter 6 should be used. It is clear that fitting all possible confounders, especially if there are many possible confounders, could lead to highly variable estimates. Some

selection is therefore advisable. An approach would be to report the range of estimates of the effect of the variable of interest given by various adjustments. If an adjusted estimate differs very little from the unadjusted estimate, there would seem to be no need for adjustment. The change-in-estimate criterion could thus be very useful but the choice as to when a change is large enough to take note of should preferably be made on clinical grounds. An odds ratio which changes from 10 to 5 may clinically be interpreted the same ("the factor has a large effect") whereas an odds ratio of 1.5 may be interpreted differently from one of 2.0. Furthermore a change may not necessarily reflect a removal of bias, but could introduce bias. To be sure that the results are not biased, changes in estimates should make sense in terms of the observed confounder patterns.

Since the procedure which minimises the error of estimation addresses the core of the estimation problem, this approach seems ideal. The choice of operating model may, however, not be clearcut, and the assessment of the presence of effect modification in the case of many possible effect modifiers may be difficult. The results based on the partial Gauss discrepancy seem to indicate that further research is necessary to determine under which conditions this approach is applicable.

REFERENCES

Achen CH (1982). Interpreting and using regression. Sage University Paper Series on Quantitative Applications in the Social Sciences, 07-029. Sage Publications, Beverley Hills.

Atkinson AC (1980). A note on the generalized information criterion for choice of a model. Biometrika 67: 413–418.

Becher H (1992). The concept of residual confounding in regression models and some applications. Statistics in Medicine 11: 1747–1758.

BMDP Statistical Software Manual (1983). University of California Press, Berkeley.

Boivin J-F and Wacholder S (1985). Conditions for confounding of the risk ratio and of the odds ratio. American Journal of Epidemiology 121: 152–158.

Breslow NE and Day NE (1980). Statistical Methods in Cancer Research. Volume 1 - The Analysis of Case-control Studies. International Agency for Research on Cancer, Lyon.

Bross IDJ (1966). Spurious effects from an extraneous variable. Journal of Chronic Diseases 19: 637-647.

Cochran WG (1968). The effectiveness of adjustment by subclassification in removing bias in observational studies. Biometrics 24: 295–313.

Collett D (1991). Modelling Binary Data. Chapman and Hall, London.

Cox DR and Snell EJ (1989). Analysis of Binary Data. Second Edition. Chapman and Hall, London.

Dales LG and Ury HK (1978). An improper use of statistical significance testing in studying covariables. International Journal of Epidemiology 7: 373–375.

Day NE, Byar DP and Green SB (1980). Overadjustment in case-control studies. American Journal of Epidemiology 112: 696–706.

Draper N and Smith H (1981). Applied Regression Analysis. Second Edition. John Wiley and Sons, New York.

Efron B (1983). Estimating the error rate of a prediction rule: improvement on cross-validation. Journal of the American Statistical Association 78: 316–331.

Efron B (1986). How biased is the apparent error rate of a prediction rule? Journal of the American Statistical Association 81: 461–470.

Efron B and Gong G (1983). A leisurely look at the bootstrap, the jackknife, and cross-validation. The American Statistician 37: 36–48.

Efron B and Tibshirani R (1993). An Introduction to the Bootstrap. Chapman and Hall, London.

Fisher L and Patil K (1974). Matching and unrelatedness. American Journal of Epidemiology 100: 347–349.

Fleiss JL (1979). Confidence intervals for the odds ratio in case-control studies: the state of the art. Journal of Chronic Diseases 32: 69–77.

Fleiss JL (1981). Statistical Methods for Rates and Proportions. Second Edition. John Wiley and Son, New York.

Fleiss JL (1986). Significance tests have a role in epidemiologic research: reactions to AM Walker. American Journal of Public Health 76: 559–560.

Gordon T (1974). Hazards in the use of logistic function with special reference to data from prospective cardiovascular studies. Journal of Chronic Diseases 27: 97–102.

Grayson DA (1987). Confounding confounding. American Journal of Epidemiology 126: 546–553.

Greenland S (1983). Tests for interaction in epidemiologic studies: a review and a study of power. Statistics in Medicine 2: 243–251.

Greenland S (1989). Modeling and variable selection in epidemiologic analysis. American Journal of Public Health 79: 340–349.

Greenland S and Neutra R (1980). Control of confounding in the assessment of medical technology. International Journal of Epidemiology 9: 361–367.

Hauck WW and Anderson S (1986). A proposal for interpreting and reporting negative studies. Statistics in Medicine 5: 203–209.

Hauck WW and Donner A (1986). Wald's test as applied to hypotheses in logit analysis. Journal of the American Statistical Association 81: 471–477.

Hauck WW and Miike R (1991). A proposal for examining and reporting stepwise regressions. Statistics in Medicine 10: 711–715.

Hauck W, Neuhaus JM, Kalbfleisch JD and Anderson S (1991). A consequence of omitted covariates when estimating odds ratios. Journal of Clinical Epidemiology 44: 77– 81.

Hemp F (1989). Neuropsychological Impact in Children Following Head Injury. Unpublished PhD thesis, University of Cape Town.

Hoffman M, Yach D, Katzenellenbogen J, Pick W and Klopper JML (1988). Mamre Community Health Project - rationale and methods. South African Medical Journal 74: 323–328.

Hosmer DW, Jovanovic B and Lemeshow S (1989). Best subsets logistic regression. Biometrics 45: 1265-1270.

Hosmer DW and Lemeshow S (1989). Applied Logistic Regression. Wiley, New York.

Lemeshow S and Hosmer DW (1984). Estimating odds ratios with categorically scaled covariates in multiple logistic regression analysis. American Journal of Epidemiology 119: 147–151.

Linhart H and Zucchini W (1986). Model Selection. John Wiley and Sons, New York.

McCullagh P and Nelder JA (1989). Generalized Linear Models. Second Edition. Chapman and Hall, London.

Mantel N and Haenszel W (1959). Statistical aspects of the analysis of data from retrospective studies of disease. Journal of National Cancer Institute 22: 719–748.

Marquardt DW and Snee RD (1975). Ridge regression in practice. American Statistician 29: 3.

Mickey RM and Greenland S (1989). The impact of confounder selection criteria on effect estimation. American Journal of Epidemiology 129: 125–137.

Miettinen OS (1974). Confounding and effect-modification. American Journal of Epidemiology 100: 350–353.

Miettinen OS (1976). Stratification by a multivariate confounder score. Americal Journal of Epidemiology 104: 609–620.

Miettinen OS and Cook EF (1981). Confounding: essence and detection. American Journal of Epidemiology 114: 593–603.

Modan M, Halkin H, Almag S, Lusky A, Eshkol A, Shefi M, Shitrit A and F Zahava (1985). Hyperinsulinaemia. A link between hypertension obesity and glucose intolerance. Journal of Clinical Investigation 75: 809–817.

Mollentze WF, Moore A, Joubert G, Oosthuysen GM, Steyn AF, Steyn K and Weich DJV (1994). Coronary heart disease risk factors in a rural and urban Orange Free State black population. South African Medical Journal, in press.

Poole C (1987). Beyond the confidence interval. American Journal of Public Health 77: 195–199.

Raftery AE (1986). Choosing models for cross-classifications. American Sociological Review 51: 145–146.

Robins JM and Greenland S (1986). The role of model selection in causal inference from nonexperimental data. American Journal of Epidemiology 123: 392–402.

Rothman KJ (1975). A pictorial representation of confounding in epidemiologic studies. Journal of Chronic Diseases 28: 101–108.

Rothman KJ (1986). Modern Epidemiology. Little, Brown and Company, Boston.

Royston P (1992). The use of cusums and other techniques in modelling continuous covariates in logistic regression. Statistics in Medicine 11: 1115–1129.

SAS Institute Inc. (1986). SUGI Supplemental Library User's Guide. Version 5 Edition. SAS Institute Inc., Cary, North Carolina.

SAS Institute Inc. (1990). SAS/STAT User's Guide, Version 6, Fourth Edition, Volume 2. SAS Institute Inc., Cary, North Carolina.

Schall R (1989). A note on variable selection in linear models. South African Statistical Journal 23: 183–194.

Schall R and Luus HG (1992). The statistical analysis of controlled clinical trials: clinically relevant hypotheses, rather than null hypotheses of no difference. Unpublished manuscript.

Schall R and Zucchini W (1990). Model selection and the estimation of odds ratios in the presence of extraneous factors. Statistics in Medicine 9: 1131–1141.

Schlesselman JJ (1978). Assessing effects of confounding variables. American Journal of Epidemiology 108: 3–8.

Schwarz G (1978). Estimating the dimension of a model. The Annals of Statistics 6: 461–464.

Spiegelhalter DJ (1986). Probabilistic prediction in patient management and clinical trials. Statistics in Medicine 5: 421-433.

Starr TB, Dalcorso RD and Levine RJ (1986). Fertility of workers. A comparison of logistic regression and indirect standardization. American Journal of Epidemiology 123: 490–498.

Stone M (1977). An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. Journal of the Royal Statistical Society B 39: 44–47.

Thompson WD (1987). Statistical criteria in the interpretation of epidemiologic data. American Journal of Public Health 77: 191–194.

Titterington DM, Murray GD, Murray LS, Spiegelhalter DJ, Skene AM, Habbema JDF, Gelpke GJ (1981). Comparison of discrimination techniques applied to a complex data set of head injured patients. Journal of the Royal Statistical Society A 144: 145–175.

Truett J, Cornfield J and Kannel W (1967). A multivariate analysis of the risk of coronary heart disease in Framingham. Journal of Chronic Diseases 20: 511-524.

Tukey JW (1991). Use of many covariates in clinical trials. International Statistical Review 59: 123–137.

Van Houwelingen JC and Le Cessie S (1990). Predictive value of statistical models. Statistics in Medicine 9: 1303–1325.

Walker AM (1986). Reporting the results of epidemiologic studies. American Journal of Public Health 76: 556–558.

Yach D and Joubert G (1988). Determinants and consequences of alcohol abuse and cigarette consumption in Mamre. South African Medical Journal 74: 348–351.